2015

Innovative assessment tasks for academic English proficiency: an integrated listening-speaking task vs. a multimedia-mediated speaking task

Hye Won Lee
Iowa State University

Innovative assessment tasks for academic English proficiency: An integrated listening-speaking task vs. a multimedia-mediated speaking task

by

Hye-Won Lee

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
Carol A. Chapelle, Major Professor
Frederick O. Lorenz
Dan Douglas
Tammy Slater
Denise Vrchota

Iowa State University

Ames, Iowa

2015

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES ix

ACKNOWLEDGEMENTS xi

ABSTRACT xiii

CHAPTER 1. INTRODUCTION 1

Statement of the Problem 1
Purpose of the Study 3
Definition of Key Terms 4
Domain Analysis 4
Authenticity 5
Task Characteristics 6
Interactionalist Approach to Construct Definition 7
Significance of the Study 8
Outline of the Dissertation 9

CHAPTER 2. SYNTHESIS OF THEORETICAL AND OPERATIONAL PERSPECTIVES 11

Approach to Test Development 12
Test Development from an Evidence-Centered Design (ECD) Perspective 13
Approaches to Domain Analysis 15
Seeking Authenticity—Attempts to Model the Target Domain 17
Authenticity in Language Learning: Widdowson's Genuineness vs. Authenticity 18
Authenticity in Language Testing: Bachman's Dual View of Authenticity 19
Evaluation of the Dual View of Authenticity 20
A Newly Interpreted Relation Between Situational and Interactional Authenticity 23
Authenticity in the 21st Century: Validation of the Simulated Target Domain 28
Academic Spoken English Discourse—The Target Domain 30
Academic Lectures 31
Academic Spoken Registers in General 34

Assessment Task Characteristics—Situations Representative of the Target Domain 42
Defining Language Ability: An Interactionalist Perspective 44
Language Knowledge, Context, and Systemic Functional Linguistics 45
Strategic Competence in Language Testing 47
Visuals in Integrated Assessment Tasks 48
The Role of Visuals in the Construct Definition 49
The Role of Visuals in Task Authenticity 52
Construct Definition of the ILST and the MMST 53
Contextual Features Simulated in the ILST and the MMST 54
Language Ability Elicited by the ILST and the MMST 56
Research Questions 59
Chapter Summary 62

CHAPTER 3. METHODOLOGY 64

Two-Part Research Design for Authenticity Analysis 64
Mixed-Methods Design: The Embedded Model 65
Qualitative Research Design with Data Transformation 66
Participants 68
Materials and Instruments 70
Multimedia-Mediated Speaking Task (MMST) 71
Integrated Listening-Speaking Task (ILST) 73
Stimulated Recall Protocol 75
Situational Authenticity Questionnaire 76
Interactional Authenticity Questionnaire 76
Data Collection Procedures 77
Task Performance 78
Stimulated Recall 79
Authenticity Questionnaires 79
Scoring 81
Data 82
Transcribed Spoken Responses on the Two Assessment Tasks 82
Assigned Task Scores 82
Transcribed Strategic Behavior Reports from Stimulated Recall 84
Authenticity Questionnaire Responses 84
Data Analysis 85
Grouping Based on the Task Scores 85
Authenticity Questionnaire Analysis 87
Systemic Functional Linguistic Analysis 88
Analysis of processes 88
Cohesion analysis 89
Strategic Competence Analysis 90
Chapter Summary 91

CHAPTER 4. RESULTS 94

Perceived Situational Authenticity of the Assessment Tasks 94
Integrated Listening-Speaking Task (ILST) 95
Multimedia-Mediated Speaking Task (MMST) 98
Statistical Comparison Between the ILST and the MMST 100
Section Summary 103
Perceived Interactional Authenticity of the Assessment Tasks 104
Integrated Listening-Speaking Task (ILST) 105
Language knowledge 105
Strategic competence 108
Multimedia-Mediated Speaking Task (MMST) 115
Language knowledge 115
Strategic competence 118
Comparison Between the ILST and the MMST 123
Language knowledge 124
Strategic competence 128
Section Summary 130
Elicited Language Knowledge Manifested in Task Performance 133
Use of Processes for Task Types: The ILST and the MMST 134
Comparisons between the ILST and the MMST 134
Comparisons between the ILST and the MMST per task score 137
Summary 140
Use of Processes for Task Versions: Biology and Business 141
Biology 142
Business 145
Summary 150
Use of Exophoric Reference for Task Types: The ILST and the MMST 151
Comparisons between the ILST and the MMST 152
Comparisons between the ILST and the MMST per task score 154
Summary 155
Section Summary 156
Strategic Competence Involved during Task Performance 158
Comparisons Between the ILST and the MMST 158
Comparisons Between the ILST and the MMST per Task Score 165
Section Summary 169
Chapter Summary 171

CHAPTER 5. CONCLUSION 175

Summary of the Main Findings 176
Implications of the Study 185
Theoretical Implications 185
Practical Implications 188
Methodological Implications 190
Limitations of the Study 192

Suggestions for Future Research 193
Conclusion 196

APPENDIX A CONTENT VISUALS FOR THE MULTIMEDIA-MEDIATED SPEAKING TASK (MMST) 197

APPENDIX B DESCRIPTION OF TASK FORMAT FOR THE MMST 199

APPENDIX C TASK SPECIFICATION FOR THE MMST 201

APPENDIX D DESCRIPTION OF TASK FORMAT FOR THE INTEGRATED LISTENING-SPEAKING TASK (ILST) 203

APPENDIX E TASK SPECIFICATIONS FOR THE ILST 205

APPENDIX F RESEARCH PROTOCOL OF STIMULATED RECALL 207

APPENDIX G SITUATIONAL AUTHENTICITY QUESTIONNAIRE 209

APPENDIX H INTERACTIONAL AUTHENTICITY QUESTIONNAIRE 210

APPENDIX I CODING SCHEME FOR STRATEGIC COMPETENCE ANALYSIS 211

REFERENCES 219

LIST OF TABLES

Page

Table 1. The domain definition inference in relation to Bachman's (1991) authenticity 30

Table 2. Contextual features simulated in the ILST and the MMST 55

Table 3. Language ability elicited by the ILST and the MMST 58

Table 4. Assignment of participants in each task 69

Table 5. Academic disciplines of the participants 70

Table 6. Overview of task assignment distribution 77

Table 7. Overview of task assignment distribution for the stimulated recall subset 78

Table 8. Descriptive statistics of the scores assigned to the ILST spoken responses 83

Table 9. Descriptive statistics of the scores assigned to the MMST spoken responses 84

Table 10. Distribution of participants across three task scores 86

Table 11. Distribution of participants from the four recruiting groups across task scores 87

Table 12. SFL analytical features 90

Table 13. Summary of data and analyses used for answering each research question 92

Table 14. Descriptive statistics for the ILST item-level ratings of situational authenticity 96

Table 15. Descriptive statistics for the MMST item-level ratings of situational authenticity 98

Table 16. Descriptive statistics for the ILST item-level ratings of interactional authenticity: Language knowledge 106

Table 17. Descriptive statistics for the ILST item-level ratings of interactional authenticity per assigned task score: Language knowledge 107

Table 18. Descriptive statistics for the ILST item-level ratings of interactional authenticity: Strategic competence 109

Table 19. Descriptive statistics for the ILST item-level ratings of interactional authenticity per assigned task score: Strategic competence 111

Table 20. Descriptive statistics for the MMST item-level ratings of interactional authenticity: Language knowledge 116

Table 21. Descriptive statistics for the MMST item-level ratings of interactional authenticity per assigned task score: Language knowledge 117

Table 22. Descriptive statistics for the MMST item-level ratings of interactional authenticity: Strategic competence 118

Table 23. Descriptive statistics for the MMST item-level ratings of interactional authenticity per assigned task score: Strategic competence 120

Table 24. Descriptive statistics for processes used in student spoken responses to the ILST and the MMST 135

Table 25. Contingency table showing the number of students who used identifying relational processes depending on the assigned assessment task 136

Table 26. Contingency table showing the number of students who used existential processes depending on the assigned assessment task 137

Table 27. Contingency table showing the number of students who used identifying relational processes depending on the assigned assessment task and task score 138

Table 28. Contingency table showing the number of students who used existential processes depending on the assigned assessment task and task score 139

Table 29. Total counts, percentages, and examples of processes used in the biology lecture 142

Table 30. Descriptive statistics for processes used in student spoken responses to the biology version of the two assessment tasks 144

Table 31. Total counts, percentages, and examples of processes used in the business lecture 146

Table 32. Descriptive statistics for processes used in student spoken responses to the business version of the two assessment tasks 148

Table 33. Descriptive statistics for pronouns as exophoric reference used in student spoken responses to the ILST and the MMST 152

Table 34. Contingency table showing the number of students who used pronouns as exophoric reference depending on the assigned assessment task 153

Table 35. Contingency table showing the number of students who used pronouns as exophoric reference depending on the assigned assessment task and task score 154

Table 36. Descriptive statistics for categories of strategy reported in strategic behavior reports 159

Table 37. Descriptive statistics for the sub-categories of communication strategies reported in strategic behavior reports 160

Table 38. Descriptive statistics for the sub-categories of cognitive strategies reported in strategic behavior reports 161

Table 39. Descriptive statistics for the sub-categories of metacognitive strategies reported in strategic behavior reports 164

Table 40. Mean reported counts of the five strategy categories per assigned task score 165

Table 41. Mean reported counts of the sub-categories of cognitive strategies per assigned task score 167

LIST OF FIGURES

Page

Figure 1. The interactionalist construct definition 45

Figure 2. The interactionalist construct definition combined with the Hallidayan model..... 46

Figure 3. The embedded model used in the study of perceived authenticity 65

Figure 4. The data transformation qualitative design used in the language knowledge study 67

Figure 5. The data transformation qualitative design used in the strategic competence study..... 67

Figure 6. A screenshot of the simulated biology lecture in the MMST 72

Figure 7. A picture of the speaker delivering the biology lecture in the ILST 74

Figure 8. An overview of research procedure and collected data for the Spring 2013 participants..... 80

Figure 9. An overview of research procedure and collected data for the Fall 2014 participants..... 81

Figure 10. The ILST mean ratings of interactional authenticity on the language knowledge items among the three score groups 108

Figure 11. The ILST mean ratings of interactional authenticity on the communication strategies items among the three score groups..... 112

Figure 12. The ILST mean ratings of interactional authenticity on the cognitive strategies items among the three score groups..... 113

Figure 13. The ILST mean ratings of interactional authenticity on the metacognitive strategies items among the three score groups 114

Figure 14. The MMST mean ratings of interactional authenticity on the language knowledge items among the three score groups 117

Figure 15. The MMST mean ratings of interactional authenticity on the communication strategies items among the three score groups..... 121

Figure 16. The MMST mean ratings of interactional authenticity on the cognitive strategies items among the three score groups 122

Figure 17. The MMST mean ratings of interactional authenticity on the metacognitive strategies items among the three score groups 123

Figure 18. The ILST and MMST mean ratings of interactional authenticity on knowledge of identifying relational processes among the three score groups 125

Figure 19. The ILST and MMST mean ratings of interactional authenticity on knowledge of existential processes among the three score groups 126

Figure 20. The ILST and MMST mean ratings of interactional authenticity on knowledge of pronouns used as exophoric reference among the three score groups 127

Figure 21. The ILST and MMST mean ratings of interactional authenticity regarding strategic competence on the overall among the three score groups 129

Figure 22. The first content visual used in the MMST 162

Figure 23. The last content visual used in the MMST 163

Figure 24. The ILST and MMST mean reported counts of the cognitive strategies among the three score groups 166

Figure 25. The ILST and MMST mean reported counts of sub-strategy of anticipating the content among the three score groups 167

Figure 26. The ILST and MMST mean reported counts of sub-strategy of using imagery among the three score groups 168

Figure 27. The ILST and MMST mean reported counts of sub-strategy of using notes among the three score groups 169

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Professor Carol Chapelle, for her guidance throughout the course of this research. Professor Chapelle has intellectually challenged me to broaden and deepen the scope of my discipline knowledge and nurtured me to be ready for conducting independent research. I feel fortunate and honored to have her as my advisor, academic role model, and lifelong mentor. I appreciate all the opportunities she has given me to grow and develop. I will cherish numerous meetings, discussions, and memories with her.

I also would like to thank my committee members, Professor Frederick Lorenz, Professor Dan Douglas, Professor Tammy Slater, and Professor Denise Vrchota, for their dedicated support. Professor Lorenz has equipped me with a strong statistical background. I especially appreciate his invaluable assistance in the quantitative analyses of this research. Professor Douglas has always encouraged me to have confidence in my research and move forward. I sincerely appreciate his advice on expanding my discussion of authenticity in Chapter 2. Thanks to him, I was able to have a more in-depth view of authenticity in language testing. I also appreciate Professor Slater's guidance on building my knowledge of systemic functional linguistics (SFL). Her expertise was a great influence on incorporating an SFL perspective into the linguistic analyses of the study. Last but not least, I thank Professor Vrchota for her suggestion about considering the effects of nonverbal communication features in future research. This idea has provided a different viewpoint to interpret the research findings.

In addition, I would like to express my sincere gratitude to Professor Roberta Vann. Professor Vann has taught me scholarly writing and helped me to edit the entire dissertation (any errors that still remain are solely my own responsibility). Every meeting with her was

delightful, encouraging, and inspiring. Her lessons have contributed to continuous improvement in my academic writing. She is a great teacher, my aunt in the United States, and my dear friend.

I would also like to thank the department faculty and my colleagues for their selfless contribution to this research. I deeply appreciate Professor Gary Ockey planting the seed of this research and willingly helping me to get permission to use the TOEFL items. I also thank my performers, Professor James Ranalli and Ms. Greta Levis, of the video lectures. Your outstanding acting made my participants think the videos were extracted from real lectures. My appreciation also goes to the three raters, Edna, Sarah, and Sinem. I thank them for spending time to rate the speech samples despite their busy graduate life. I also thank Monica for recording the aural instructions of the tasks and Russ for sharing his expertise in the use of visuals in second language assessments. Lastly, I feel thankful for all ESL instructors who allowed me to visit their classes and fully supported me in participant recruitment.

I want to also offer my appreciation to those who were willing to participate in my study without whom, this dissertation would not have been possible.

I thank my friends for making my time at Iowa State University a wonderful experience. Yooree, Hee Sung, Shinyoung, Hyejin, Yongkook, and Mijin— thank you for your advice, sympathy, and emotional support all the time. I also thank Zhi, who was a good companion along the journey of dissertation writing.

Finally, I thank my parents for their encouragement and my husband for his years of patience, respect and love. They were my motivator to persevere in this accomplishment.

This dissertation research was funded by the Small Grant for Doctoral Research in Second Language Assessment from Educational Testing Service.

ABSTRACT

Language testing researchers and test developers have long preferred to use integrated tasks due to their authentic nature. However, such presumed authenticity is taken for granted without empirical justification. Therefore, in this study, I examined and compared the authenticity of two integrated task types, a speaking task following an audio lecture on an academic topic and a multimedia-mediated speaking task (MMST)—a similar task to the integrated listening-speaking task (ILST), but with a *video* lecture as its stimulus.

First, I used a questionnaire to analyze student perception of 1) *situational authenticity* regarding the task characteristics of the two assessment tasks compared with those of a target language use (TLU) task in the academic context and 2) *interactional authenticity* regarding the involvement of language ability in accomplishing the two assessment tasks. Second, as another way to analyze the interactional authenticity, I investigated if language ability, consisting of language knowledge and strategic competence, was used as defined in the construct of the ILST and the MMST. For the language knowledge analysis, I collected student speech samples and analyzed them from a systemic functional linguistic (SFL) perspective. For the strategic competence analysis, I conducted stimulated recall interviews to obtain students' strategic behavior reports. Ninety-three international undergraduate and graduate students participated in the study.

Students perceived a similar degree of situational and interactional authenticity of both the ILST and the MMST overall. In addition, both the ILST and the MMST functioned well in eliciting some SFL features of the target language knowledge. However, some other SFL features were elicited more successfully by the MMST than by the ILST. For strategic

competence, students in both the ILST and the MMST groups reported a frequent use of communication, cognitive, and metacognitive strategies. However, students in the MMST group reported a more frequent use of cognitive strategies than those in the ILST group. These findings lead to an enhanced understanding of how the two types of integrated assessment tasks work on student performance and can provide empirical evidence for the domain definition inference of a validity argument for an academic English proficiency test.

CHAPTER 1. INTRODUCTION

1.1. Statement of the Problem

It has been a widely accepted practice in test development and administration to assess one's language proficiency with a test divided into skill sub-sections (e.g., listening, reading, speaking, writing). For instance, highly popular English proficiency tests such as the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) consist of four skill-sections and report scores by each skill area. Many test score users share this view of language ability as conceptualized by separate skills and prefer to receive scores in each skill (Jamieson, Eignor, Grabe, & Kunnan, 2008).

However, a recent approach to language testing largely acknowledges that much "real world" communication includes the integration of two or more skills (Frost, Elder, & Wigglesworth, 2012; Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000). This perspective has been reflected in test design with the adding of integrated tasks. Test takers who carry out an integrated task are provided with the input, and this stimulus material becomes the basis of test takers' performance on the task in a context that simulates real-life language use (Lewkowicz, 1997). This type of assessment task is particularly favorable for assessing academic language ability (Plakans, 2013.)

Research studies on integrated tasks in second language assessment have placed much focus on writing assessments (e.g., Cumming et al., 2006; Ohkubo, 2009; Plakans, 2009; Plakans & Gebril, 2012). For instance, a body of empirical research on integrated writing assessments appears in the latest special issue of *Language Assessment Quarterly* (2013), a well-recognized journal in the field of language testing.

However, compared to the great interest in the study of integrated tasks in writing assessment, less research has explored the use of integrated tasks in measuring speaking ability (e.g., Brown, Iwashita, & McNamara, 2005; Huang, 2010; Inoue, 2009; Iwashita, Brown, McNamara, & O'Hagan, 2008; Lee, 2006; Sawaki, Stricker, & Oranje, 2009), which is the focus of this dissertation study. These few published studies piloted prototype tests in which integrated speaking tasks were only one small portion of the test tasks. The influence of integrated tasks on test performance and score interpretation was one of the research concerns, but not the main focus of the studies.

Frost, Elder, and Wigglesworth (2012) were the first to publish a publicly accessible research study that focused on an integrated speaking task. The researchers investigated test takers' use of the input material in their speaking performance on a listening-then-speaking summary task. They examined whether this content-related feature of task completion was reflected in the task rating scale and whether scores assessed using this rating scale separated out test takers according to the quality of their oral performance. Their study findings shed light on a possibility of incorporating content-related aspects of integrated speaking task performance into rating scales and evaluation. In addition, they suggested the importance of input material use on second language speaking construct measured by integrated tasks, which the current dissertation study also supports.

Despite the limited research on integrated tasks, languages testing researchers and test developers have long preferred to use these due to their authentic nature (Plakans, 2013; Weir, 1990). Authenticity in language testing means resemblance between real-life tasks that an assessment task intends to simulate and the corresponding assessment task that involves elicited test-taker performance. Independent tasks, a traditional task type designed to assess a

productive language skill (e.g., responding orally to a prompt that asks about test takers' pleasant and memorable school event), in academic language proficiency tests can assess an individual's ability to produce language based only on his/her prior topical knowledge. In this respect, a scope of spoken or written samples that can be elicited is limited to what knowledge/experience test takers bring to the test. On the other hand, integrated tasks, where input materials are provided for the follow-up language production (e.g., explaining orally about an academic concept test takers learn from a biology input lecture), can allow test takers to talk or write about more various kinds of topics, not restricted to their prior experience or common knowledge; thus, an academic language test that includes integrated tasks has potential for evaluating a kind of language ability essential in academic situations. What test takers are asked to perform in integrated tasks is closer to what they would do in real-life academic contexts, in other words, more authentic.

However, such a presumed authentic nature of currently used integrated tasks is taken for granted without empirical justification. It is unknown 1) how much the characteristics of integrated tasks correspond to those of the target language use (TLU) tasks and 2) how much the linguistic knowledge, processes, and strategies required for successful task completion correspond to those for completing the TLU tasks (Hulstijn, Coopmans, van Hout, & Bos, 2003).

1.2. Purpose of the Study

In this dissertation, I chose to examine the authenticity of one of the integrated task types, a speaking task following an audio lecture on an academic topic. I analyzed its task characteristics and construct with the intention of comparing them with those of a TLU task

in an academic context. Authenticity questionnaires were used to examine how stakeholders of the TLU situation—students in the context of current study—perceived the extent to which the integrated task is close to the TLU task, a task of explaining aspects of an academic concept. Students' spoken responses to the task and reports during a stimulated recall interview were additional sources of data to investigate how students' language ability, consisting of language knowledge and strategic competence, was actually involved in task performance.

In addition, I developed a multimedia-mediated speaking task (MMST)—a similar task to the integrated listening-speaking task (ILST) but with a *video* lecture as its stimulus. With this task, I sought to achieve the following two goals: 1) to test the possibility of multimodal input in integrated tasks and 2) to increase authenticity of response as well as input in using such tasks (Chapelle & Douglas, 2006; Douglas & Hegelheimer, 2007). I examined the MMST in terms of its comparability with the ILST as well as the TLU task.

1.3. Definition of Key Terms

The terms and concepts introduced below are central to understanding the goal and design of the current study, and defined in the following sub-sections.

1.3.1. Domain Analysis

The target domain, to which interpretations about examinees' language ability elicited by the ILST and the MMST intend to generalize, is an academic spoken English context, and it is important to learn about the characteristics of this domain and use learned information in designing and developing the assessment tasks. An Evidence-Centered Design (ECD)

perspective, a test design framework widely known in the field of educational measurement, provides a conceptually systematic means to connect information gathered about the target domain to the aspect of ability to be assessed (Mislevy, Steinberg & Almond, 2003; Mislevy & Haertel, 2006b; Mislevy, 2011), and the first ECD layer, *domain analysis*, yields "substantive information about the domain to be assessed" and helps "understand the knowledge people use in a domain, the representational forms, characteristics of good work, and features of situations that evoke the use of valued knowledge, procedures, and strategies" (Mislevy & Haertel, 2006b, p. 7). In the current study, relevant literature on the academic spoken English discourse was analyzed to learn about the domain the ILST and the MMST attempted to simulate (refer to Section 2.3).

1.3.2. Authenticity

Developing assessment tasks aligned with the target domain is a key element in educational measurement practices following the ECD perspective. In the field of second language testing, reflecting the features of the target domain in assessment task development has also been long recognized under the concept of *authenticity*. Creation of "authentic" assessment tasks is the way in which test developers pursue simulating the target language use domain in their design process. Bachman (1991) viewed authenticity as composed of two types: *situational authenticity* and *interactional authenticity*. *Situational authenticity* highlights a relationship between the characteristics of an assessment task and those of the corresponding real-life situation, whereas *interactional authenticity* refers to the involvement of a test taker's language ability in completing an assessment task.

Authenticity can be explicitly related to argument-based validation (Chapelle & Lee, 2013). The domain definition inference, which incorporates Mislevy's ECD perspective, links performances in the TLU domain to the observations of performance on assessment tasks. In order to warrant this connection, the following statement needs to be justified: Observations of performance on assessment tasks reveal relevant knowledge, skills, and ability in situations representative of those in the TLU domain. Domain analysis, expert judgment and empirical analysis of task simulation can substantiate the relevance of assessment tasks to TLU tasks. These analyses examine the representativeness of assessment tasks and the relevance of language ability in accomplishing those tasks. In this regard, what is pursued in the domain definition inference is parallel to Bachman's (1991) situational authenticity and interactional authenticity.

In this dissertation, situational and interactional authenticity of the ILST and the MMST was investigated. The research results provided empirical evidence to support or refute the domain definition inference of a validity argument for an academic English proficiency test that contains the types of integrated tasks featured in this study.

1.3.3. Task Characteristics

In this research, an assessment task was considered a vehicle for simulating an authentic situation comparable to the target domain, and therefore, during the task design phase, it was important to detail the characteristics of task that contribute to high situational authenticity. In language testing, Bachman and Palmer's (1996; 2010) task characteristics framework is widely used for describing aspects of tasks, which was also used in the current study. Bachman and Palmer's framework is comprised of five facets: 1) setting, 2) rubric, 3)

input, 4) expected response, and 5) relationship between input and expected response. This task characteristic framework provides a list of task features that developers must consider simulating the target domain to the extent possible, and a system to incorporate the domain knowledge into task development. Aspects featured in the framework can also be used to evaluate how successful an assessment task models the target language use task, namely the degree of situational authenticity.

1.3.4. Interactionalist Approach to Construct Definition

Domain analysis informs the description of both task characteristics and the language ability a task requires. Language ability is a highly abstract concept that does not have a readily available tool for its description, but for assessment purposes, it needs to be defined in concrete terms. We call this defined ability a *construct*, which "provides the basis for a given assessment or assessment task and for interpreting scores derived from this task" (Bachman & Palmer, 2010, p. 43). In the field of measurement, three fundamental theoretical perspectives to construct definition exist: trait, behaviorist, and interactionalist (Chapelle, 1998). In the current study, I used the interactionalist perspective to explicate the construct measured by the two assessment tasks, the ILST and the MMST. This perspective is consistent with current theory in applied linguistics (Bachman & Cohen, 1998; Bachman & Palmer, 2010), and useful for describing language ability, which is composed of both language knowledge and strategic competence interacting in the context of language use.

From the perspective of interactionalist construct definition, the context of language use influences the linguistic choices a language user can make during task performance, which can be described by Halliday's (1978) model of language in context. When task

performers produce a meaningful sample of linguistic performance (a text) in a given task, they make choices from their knowledge of the language system, in order to meet the demands of the context in which they create the text.

Strategic competence is another one of the individual factors in an interactionalist construct definition. Strategic competence is defined as a set of "higher-order metacognitive strategies that provide a management function in language use" (Bachman & Palmer, 2010, p. 48), and controls the integration of the individual attributes (e.g., language knowledge) when assessing language use situations. Three general areas in which metacognitive strategies operate are: 1) goal setting (deciding what one is going to do), 2) assessing/appraising (taking stock of what is needed, what one has to work with, and how well one has done), and 3) planning (deciding how to use what one has) (Bachman & Palmer, 1996; 2010). Since metacognitive strategies are mental processes, they are not directly observable and therefore need to be inferred by researchers through introspective methods. Therefore, in the current study, I looked for evidence of strategic competence as "*reported* actions and thought processes" from the participants, following the definition used in a recent strategic behavior study in language testing conducted by Swain, Huang, Barkaoui, and Brooks (2009) (p. 2, emphasis in original).

1.4. Significance of the Study

This dissertation is significant for the following reasons. First, the study provides empirical evidence to evaluate the apparent authenticity of integrated assessment tasks from Bachman's (1991) dual view of authenticity. Second, methodologically, this study contributes to the language-testing field in that it provides an extensive analysis of language

ability involved in task performance based on the interactionalist approach. Third, the current study analyzed test takers' spoken responses from a systemic functional linguistic (SFL) perspective, and this linguistic analysis adds to a repertoire of analytic tools used for studying language samples from second language assessments. Lastly, this dissertation study pilots an innovative type of integrated tasks, the multimedia-mediated speaking task (MMST) I have developed for the study, and suggests the reasonable possibility of using visuals in integrated assessment tasks.

Overall, this dissertation should provide a fuller understanding of the nature of integrated assessment tasks that require using oral communication skills, and better inform task selection in the course of test development. Further, it contributes to interpreting the meaning of performance from this type of task, and eventually, serves one type of support in validity arguments for an academic English proficiency test that contains similar types of integrated tasks.

1.5. Outline of the Dissertation

This dissertation consists of five chapters. The current chapter introduced the problem this study intends to resolve, the main goals of the study, the definitions of key terms and concepts of the study, and the significance it provides. The second chapter consists of eight main parts. The first part discusses an approach to test development called evidence-centered design (ECD), specifically its first step, domain analysis. Since an attempt to model the target domain in test development and validation is conceptually equivalent to seeking to construct authentic tasks, the second part provides an overview of authenticity in language learning and testing, showing how the conception of this notion has evolved, and presents the authenticity

framework for this dissertation. For understanding the characteristics of the target domain of tasks featured in the current study, the third part reviews previous studies on academic spoken English discourse. The fourth part explains the task characteristic framework the current study used to describe the context the assessment tasks intend to simulate. The fifth part deals with a conceptual discussion of the interactionalist perspective on defining language ability. This perspective relates to the construct definition of the two assessment tasks used in this dissertation, the ILST and the MMST. The sixth part demonstrates grounds for using visuals in integrated assessment tasks, and aims to justify the use of visual information in the integrated task I developed, the MMST. The seventh part describes the construct of the ILST and the MMST, reflecting the theoretical underpinning discussed in the previous parts. The last part of the chapter articulates the research questions the current study examined.

The third chapter first describes the research design the current study followed, then, the characteristics of study participants, and third, the materials and instruments used to elicit and obtain data. The fourth part explains the procedures for participant recruitment and data collection, followed by the description of collected data. The last part of Chapter three details how the data were analyzed. The fourth chapter presents the findings of the data analyses for answering each of the research questions and discusses how the findings can be interpreted. The last chapter begins with a summary of the main findings, and second, addresses theoretical, practical, and methodological implications. The third part acknowledges limitations of the study, and the fourth part suggests possibilities for future research. The chapter ends with a conclusion.

CHAPTER 2. SYNTHESIS OF THEORETICAL AND OPERATIONAL PERSPECTIVES

The purpose of my dissertation is to evaluate how authentic two types of integrated speaking tasks, 1) an integrated listening-speaking task (ILST) and 2) a multimedia-mediated speaking task are, and in other words, to what extent the assessment tasks correspond to the target language use tasks in the view of stakeholders. In addition, the study seeks to examine what components of language ability are measured by the tasks. The purpose of this chapter is to discuss theoretical and operational perspectives that form the basis of my dissertation study. Theoretical perspectives include a synthesis of conceptual aspects I incorporated into my task development and research framework, and operational perspectives consist of considerations of specific features of my tasks, and the definition of the operational construct.

The first section, dealing with theoretical perspectives, discusses the approach to test development, specifically an Evidence-Centered Design (ECD) approach. Analyzing the target domain from an ECD perspective was a basis for deciding how to develop my integrated tasks and what should be measured by them. The second section provides a thorough review and new conception of authenticity I sought in order to attempt to model the target domain. These first two sections provide a theoretical framework of task development and authenticity analysis and guide the rest of the chapter. Third, relevant literature on academic spoken English discourse, the target domain of my tasks, is reviewed to learn what constitutes this domain. This section concerns the first step of task development, domain analysis.

Next, a framework of assessment task characteristics is explained, which was used as a tool to describe the domain simulated in my tasks and an analytical structure to evaluate situational authenticity of my tasks. Fifth, one of theoretical ways of perceiving a definition of language ability, an interactionalist perspective, is introduced. This approach is the basis for conceptualizing the definition of my task construct and evaluating interactional authenticity. The fourth and fifth sections provide guidelines of defining task characteristics and language ability. In Section 2.6, dealing with operational perspectives, visuals in integrated assessment tasks are discussed, as this is a unique characteristic of one of my tasks, the MMST. The seventh section presents the construct definition of the ILST and the MMST, following an interactionalist perspective. The sixth and seventh sections are operational, the former discussing a unique characteristic of one of my tasks, and the latter presenting the construct definition of my tasks. Lastly, research questions of the current study are presented, followed by a chapter summary.

2.1. Approach to Test Development

Test developers seek to create assessment tasks that can provide useful information about the ability of test takers. Some language researchers suggest that such tasks, in the context of language assessments, should correspond to target language use (TLU) tasks, whose characteristics need to be the basis of assessment tasks. In addition, assessment tasks need to elicit performance that will allow for inferences to be made about the construct to be assessed so that the interpretations of task performance can be meaningful for their intended purpose.

A systematic development process is needed to produce quality tasks, and score interpretation from the tasks can be generalized to a wider, real-life context. Bachman and Palmer (1996; 2010) highlight that such generalizability is built upon consideration of the target language use (TLU) domain during test development. They advise that test developers first should identify and describe the TLU domain that is relevant for score interpretation. Among the infinite tasks in the target domain, those that can serve as a basis for developing language assessment tasks need to be selected by test developers, and the characteristics of the selected TLU tasks need to be described. This domain analysis guides the rest of the task development procedure, especially in terms of delineating the characteristics of the target assessment tasks and ultimately writing task specifications. Close correspondence between the target domain and the context simulated in a test at this stage of development is important for arguing the generalizability of test performance to a broader context, allowing test scores to be interpreted more meaningfully, and not confined to a test setting.

2.1.1. Test Development from an Evidence-Centered Design (ECD) Perspective

While Bachman and Palmer (1996; 2010) emphasize the need for and analysis of the target domain as part of test development, they are less explicit about the nature of such an analysis or the use of the obtained knowledge in contributing to a definition of the assessed ability. An Evidence-Centered Design (ECD) perspective, a test design framework widely known in the field of educational measurement, provides a conceptually more systematic means to connect information gathered about the target domain to the aspect of ability to be assessed (Mislevy, 2011; Mislevy & Haertel, 2006b; Mislevy, Steinberg & Almond, 2003). This framework helps make explicit how domain research is "embodied in the elements … of

an assessment" including a construct to be measured in an assessment (Mislevy et al., 2014, p. 128).

The first two ECD layers, *domain analysis* and *domain modeling*, particularly manifest a close connection between identified domain characteristics and an assessment construct. First, *domain analysis* yields "substantive information about the domain to be assessed" and helps "understand the knowledge people use in a domain, the representational forms, characteristics of good work, and features of situations that evoke the use of valued knowledge, procedures, and strategies" (Mislevy & Haertel, 2006b, p. 7). At this stage in assessment design, developers must concern themselves with important features of the target domain regarding central abilities requested in the domain, representations of such abilities, and contexts eliciting the incidences of represented abilities. In the current study, relevant literature on the academic spoken English discourse was analyzed to learn about the domain the integrated assessment tasks attempt to simulate (refer to Section 2.3).

Domain modeling is "a transition from the substantive, specialized compendium of knowledge about the target domain to forms that guide the building of the assessment machinery" (Mislevy & Haertel, 2006a, p. 7). At this stage, test developers conceptualize information gathered from *domain analysis* and describe relationships among what has been collected. The collected information can be translated into the following three components: 1) the focal knowledge, skills, or abilities, namely constructs, 2) aspects of what test takers do that can provide evidence about the primary constructs, and 3) features of task characteristics that can evoke the desired evidence. Centered around the constructs, attributes of task characteristics and evidence about what test takers do are organized in a way to guide the rest of test development. It is critical to specify constructs and task characteristics,

derived from a thorough domain analysis, for designing assessment tasks that can enable stakeholders to make informed interpretations of test takers' abilities in the targeted context. The current study utilized information learned from the analysis of relevant literature for defining a construct to be assessed, and developing tasks that elicit performance to be used for assessing the defined construct, including characterizing the expected examinee performance.

2.1.2. Approaches to Domain Analysis

Domain analysis is the cornerstone of the subsequent test design procedure in the ECD framework, and involves a detailed study of target abilities and situations from any number of resources such as "cognitive research, available curricula, professional practice, ethnographic studies, expert input, standards and current testing practices" (Mislevy, 2011, p. 8). From a brief examination of some documented projects using the ECD framework and materials particularly on *domain analysis*, I discovered that the following three approaches have been frequently utilized to conduct *domain analysis*: 1) reviewing existing literature regarding the target domain, 2) referring to relevant standards, and 3) consulting domain experts and/or analyzing domain practices.

First, the target domain can be analyzed through reviewing relevant literature. For instance, Mislevy and Haertel (2006a), in one of their Principled Assessment Designs for Inquiry (PADI) Projects, referred to numerous research articles, books, and conference presentations on science inquiry, and applied learned knowledge to produce assessment tasks that evaluate students' ways of scientific thinking. Such a literature review allowed the developers to identify key knowledge, skills, and abilities in the context of science. Another

example is Mislevy et al.'s (2014) game-based assessment project, which drew on previous research in science education to help them understand the core abilities of scientific reasoning.

Relevant standards can be another useful source of *domain analysis*. Webb (2006) suggested referring to standards such as the National Science Education Standards to describe the subject-area content to be assessed in student achievement tests. Mislevy et al. (2014) consulted another set of science education standards, the Next Generation Science Standards, to identify core disciplinary ideas and concepts that need to be included in science assessment.

Consulting domain experts and/or analyzing real-life practices are additional ways to examine the target domain. Steinberg et al. (2003) engaged a set of high school biology teachers and university professors in several biology sub-areas during the process of their prototype development for the Biomass project. The goal of the project was to create an assessment that evaluates standards-based learning in biology, and the developers relied on knowledge of the subject-matter experts as well as collections of science standards in deciding what to be assessed and how. Practice analysis, that is, the study of task characteristics required for successful accomplishment of particular jobs, can also be regarded as one of the domain analysis methods, especially for credential testing (Raymond & Neustel, 2006). In this analysis, the kinds of knowledge, skills, and abilities test users need to draw inferences are identified to test the qualifications of the applicants.

2.2. Seeking Authenticity—Attempts to Model the Target Domain

Developing assessment tasks aligned with the target domain is a key element in educational measurement practices following the ECD perspective. In the field of second language learning and testing, reflecting the features of the target domain in learning and assessment task development has also been long recognized under the concept of *authenticity*. Creating "authentic" learning activities and tests is the way in which materials developers pursue simulating the target language use domain in their design process. By using authentic materials, language teachers seek to provide opportunities for their students to practice language use activities that they will encounter outside of the classroom. By including authentic tasks in a language assessment, test developers intend to evaluate test takers' language ability to fulfill corresponding real-life tasks. In other words, pursuing *authenticity* is another way to describe attempts to model the target domain in developing language use tasks.

In this sub-section of the chapter, I will discuss how the concept of authenticity has historically evolved first in language learning, particularly Widdowson's view on genuineness and authenticity, and second in language testing, Bachman's dual view of authenticity. Next, I will explain how authenticity can be evaluated, including a review of previous authenticity studies. Lastly, I will propose a new way of viewing the relationship between Bachman's dual view of authenticity from the strict sense of Widdowson's, and present how this perspective is connected to the current conception of validating the target domain modeling in an assessment.

2.2.1. Authenticity in Language Learning: Widdowson's Genuineness vs. Authenticity

Communicative language teaching (CLT), a growing teaching approach in the late 70s, objected to the emphasis on decontextualized language structure and rules in language instruction, and embraced the concept of "communicative competence" (Canale, 1983; Canale & Swain, 1980; Hymes, 1967, 1972) and the importance of an ability to *use* the target language, not simply to *know* about the language. Here, the role of "authentic" language use became important. In CLT, teachers were encouraged to engage students in the "authentic … use of language for meaningful purposes" (Brown, 2000, p. 166). Classroom activities in communicative language teaching emphasize communication through interaction in the target language such as role plays and interviews.

Widdowson (1978; 1979; 1983), one of the CLT proponents, distinguished *authenticity* from *genuineness*: "*Genuineness* is a characteristic of the passage itself and is an absolute quality. *Authenticity* is a characteristic of the relationship between the passage and the reader and has to do with appropriate response" (Widdowson, 1978, p. 80, emphasis added). Though Widdowson was referring only to reading, due to the relative weight on the written medium in second language (L2) teaching at the time, the relationship between genuineness and authenticity was later extended to a broader teaching context, including both oral and written language (Douglas, 2000). In this expanded sense, genuineness is for "the actual spoken or written texts produced by the users" and authenticity for "activities or processes associated with instances of language use" (Douglas, 2000, p. 17). In other words, genuineness is a quality of oral and written texts that serve as input to students, whereas authenticity is an attribute assigned to the texts by their audience.

In Widdowson's view, using an intact text in learning tasks does not guarantee authenticity, because the text is taken out of its original context and placed into a new one, where it may or may not be seen as authentic. Only when this new context generates students' involvement in comprehending the author's intentions and conventions as the original audience of the text would do, can the texts produce an authentic response. Context where an authentic interaction between the text and its audience is created can contribute to a high degree of authenticity.

2.2.2. Authenticity in Language Testing: Bachman's Dual View of Authenticity

A communicative approach was also prevalent in the late 70s to early 80s in language testing (e.g., Alderson, 1981; Canale 1984; Carroll, 1980; Harrison 1983; Morrow; 1979). In this communicative approach to language testing, using "authentic" test settings and materials was integral, and consequently, the notion of *authenticity* became one of the centers of attention among language testing professionals. Since then, authenticity has been one of the major considerations in test design and validation (Bachman, 1990; Spolsky, 1985), and its definition has evolved as the profession has developed. Bachman (1990), in his influential language testing book, *Fundamental Considerations in Language Testing*, defined and operationalized authenticity from two approaches, the "real-life" approach (p. 301) and the "interactional/ability" approach (p. 302). In this book, authenticity was viewed as a unitary notion but approached from two perspectives. This viewpoint soon led to the definition of two types of authenticity: *situational authenticity* and *interactional authenticity* (Bachman, 1991). *Situational authenticity* highlights a relationship between the characteristics of an assessment task and those of the corresponding real-life situation, whereas *interactional*

authenticity refers to the involvement of a test taker's language ability in completing an assessment task.

This dual view of authenticity is one of the key factors to distinguish language tests for relatively specific purposes from those for relatively general purposes (Douglas, 2000). If an assessment task closely corresponds to tasks in a specific setting such as a situation of ordering food at a restaurant, performance on that task can be assumed to reflect one's language ability in real life for that specific purpose. The interwoven characteristics among the target language use (TLU) situation, assessment tasks, and test takers' language ability are echoed in assessing language for specific purposes (LSP) in the sense that an LSP assessment task simulates a particular TLU situation, and elicits and assesses language performance that captures a language ability to be expectedly used in the target situation.

Subsequently, Bachman and Palmer (1996) extended the scope of the "interactional/ability" side of authenticity to include more individual characteristics besides language ability in the definition (p. 42). It acquired a new name, "interactiveness," and is defined in that book as a distinct entity from authenticity. However, the importance of the interactional aspect in test development and use is still in line with Bachman's (1991) earlier notion of authenticity.

2.2.3. Evaluation of the Dual View of Authenticity

When it comes to evaluating authenticity, the following questions are generally addressed:

1) To what extent do the characteristics of the assessment task correspond to those of tasks in the target language use (TLU) situation?

2) To what degree do the stakeholders such as the test developer and users consider the assessment task to be situationally authentic overall?
3) To what degree do the stakeholders consider the involvement of language use ability in the assessment task to be high?

(Adapted from Bachman & Palmer, 1996, pp. 263-264)

Bachman and Palmer's (2010) framework, consisting of five aspects of language task characteristics—the setting, rubric, input, expected response, and relationship between input and response, is useful in comparing the characteristics of TLU and assessment tasks. It can be used for creating an instrument such as a questionnaire to use in research intending to provide empirical data consisting of situational authenticity. The perceptions of test users measured through such research instruments can support a claim about an overall degree of authenticity in terms of context simulated by tests.

Bachman and Palmer's (2010) framework provides a springboard for specifying task characteristics. However, it does not adequately address all aspects of tasks, and determining the "critical" features of TLU tasks seems to "require judgment" (Lewkowicz, 2000, p. 50). In this sense, the estimates of authenticity have largely depended on the opinions of stakeholders such as language testing experts, test takers, and potential participants of the TLU task.

In an attempt to develop authentic input materials for an LSP test used in the US Law Enforcement Agencies (LEAs), Wu and Stansfield (2001) asked for comments and criticism from the LEA linguists, who were the real world task performers. This test development process aimed for a structured and iterative process to check authenticity in the course of task

development. The researchers sought a close correspondence between their tasks and the TLU tasks using the field experts' judgment.

Another study that investigated the representativeness of the assessment task, Cumming, Grant, Mulcahy-Ernt, and Powers (2005), asked college instructors to respond to a questionnaire item on how well the speaking and writing prototype tasks for the new Test of English as a Foreign Language (TOEFL) represented the domain of an English-medium university in North America. The degree of "content validity[1]," which is equivalent to Bachman's (1991) situational authenticity, was evaluated by the perceptions of the domain experts (Cumming et al., 2005, p. 13).

Other empirical studies on authenticity examined its interactional side. Lumley and Brown (1998) elicited comments from nurses and analyzed the discourse that test takers and interlocutors co-produced in an LSP test for health professionals. After taking the test, these medical experts reported on the inauthenticity of interactions elicited from the assessment tasks. In another study, Lewkowicz (2000) asked a group of Cantonese-speaking freshmen to respond to a questionnaire on how they perceived the degree to which the academic English tests they had completed reflected their academic language ability in real life. Besides using interviews and questionnaires to research interactional authenticity, Spence-Brown (2001) included reports from stimulated recall interviews about test takers' engagement with the task.

These research studies have dealt with one or more aspects of the multi-faceted concept of authenticity, but not from a comprehensive viewpoint. Each contributes to a partial understanding of how authenticity can be studied in addition to providing research

[1] "Content validity" in Cumming, Grant, Mulcahy-Ernt, and Powers (2005) refers to how well the assessment tasks represent the TLU domain.

results that are relevant to a particular testing context. Together, they also support the need identified by Lewkowicz (2000) for a systematic investigation of authenticity. The first such comprehensive investigation was an extensive study by Liu (2005). Drawn from a synthesis of multiple theoretical frameworks, Liu devised a thorough evaluation tool to assess the authenticity of tasks in a computer-assisted English learning program. She reviewed relevant literature and analyzed the TLU situation. Followed by this domain analysis, she investigated both the correspondence of the learning tasks to the TLU tasks and the degree of learner involvement in language production to obtain an evaluation of both situational and interactional authenticity. This evaluation resulted from various kinds of data collection, including the analysis of actual speech samples produced by research participants.

2.2.4. A Newly Interpreted Relation Between Situational and Interactional Authenticity

According to Douglas (2000) and Lewkowicz (2000), Bachman's (1991) dual view of authenticity was heavily influenced by Widdowson and his seminal work (1978; 1979; 1983). If Widdowson's view on genuineness and authenticity is adopted in language testing, *genuineness* will be a characteristic of assessed target language use (TLU) tasks performed in the TLU context, and by definition, the tasks will be genuine. However, as soon as these tasks are detached from the TLU context and made part of a language assessment, they will be no longer genuine. Now the focus shifts from genuineness to authenticity. "*Authentic*" assessment tasks should first simulate the TLU tasks as closely as possible and second, engage task performers to the best possible extent as if they were performing the tasks in the TLU context. The first aspect, simulating the TLU tasks, corresponds to Bachman's (1991) *situational authenticity*, defined as "the perceived relevance of the [task] characteristics to the

features of a specific target language use situation" (p. 690), and the second aspect, engaging task performers, to *interactional authenticity*, "a function of the extent and type of involvement of task taker's language ability in accomplishing a test task" (p. 691).

Situational authenticity seeks an authentic reproduction of "genuine" TLU tasks in the testing context. A high degree of this authenticity type should be a precondition for interactional authenticity, which Bachman (1991) clarifies as "essentially Widdowson's (1978) definition of authenticity" (p. 691). As Widdowson (1978) emphasizes the reader's "appropriate response" to the passage in his definition of authenticity (p. 80), Bachman's interactional authenticity values the task taker's appropriate *response* to a test task. The task taker's appropriate response indicates the use of the performers' language ability appropriate to the context established in a test task.

Based on the discussion of authenticity so far, task authenticity should be considered from perspectives of both task contexts and task takers' engagement with those contexts, and a higher degree of comparability between the TLU tasks and the assessment tasks must be a necessary condition for a successful elicitation of expected language ability, namely, high interactional authenticity. In other words, if an assessment task lacks situational authenticity, task performers will be less likely to engage with the anticipated processes of accomplishing a task.

Interactional authenticity must be dependent on the level of situational authenticity if we strictly follow Widdowson's notion of authenticity. Widdowson's authenticity, which Bachman (1991) termed *interactional* authenticity, assumes that the process of the reader's response to the passage is not "contrived in a contextual vacuum" (Widdowson, 1979, p. 170). For Widdowson, by the time the passage is detached from its original context, it loses

its genuineness. The same occurs when the TLU task is transferred to the assessment context. We cannot avoid the contrived nature of reading a "genuine" passage in the learning context, and the same is true for accomplishing an assessment task that simulates the corresponding TLU task[2] in the testing context. However, this contrived activity of reading or accomplishing a task is not completely context-less. Widdowson (1979) underscores the integral role of a teaching method that promotes a context very similar to real life, and this method "is only adequate to the extent that it [creates an authentic response]" (p. 171). An adequate context should reproduce "situationally authentic" circumstances in Bachman's (1991) term (p. 690), and also generate an appropriate response, a highly authentic interaction between the passage and the reader or between the tasks and the task performers, namely high interactional authenticity. In this sense, Bachman's (1991) two types of authenticity work in tandem.

However, although Bachman's (1991) view on authenticity succeeded Widdowson's, the relationships between Bachman's situational and interactional authenticity shown in his Figure 4 (p. 693) and the associated examples (pp. 692-694) appear to consider the two authenticity types completely independent. In Bachman's view, tasks characterized as having high situational and low interactional authenticity or low situational and high interactional authenticity can exist. His first example (high situational and low interactional authenticity) is a screening task where an applicant for a typist position is asked to type from a handwritten document. The context created by this task is highly relevant to the real-life context for typists. It is a "genuine" task in the typist world, and therefore, a situationally highly authentic assessment task. However, since this task involves only "mechanical control

[2] In the case of a reading passage, it is easy to extract an intact text and put it in another context. On the other hand, when it comes to a language use task, one in the testing context has to be always a simulation of the TLU task because it is impossible to replicate the real-life task exactly the same in a test situation.

of English," just typing what is visually recognized without much understanding of its underlying meaning (Bachman, 1991, p. 692), Bachman interprets this as being interactionally low in authenticity. While it is true that the level of language ability involved in this task is not high, from my perspective, it is an absolutely *appropriate response* to the given context that a typist in real life would also be expected to do. Thus, from this perspective, it is both situationally and interactionally authentic. It is not the case that this task is useful in spite of its low interactional authenticity, but it is useful because it is highly authentic in terms of both situations and interactions.

Bachman's second example (low situational and high interactional authenticity) uses the same testing condition as the example above, a screening task for typists. This time, an interview task is used to select a typist. Bachman evaluates this task as low in situational authenticity because having a conversation in the target language is not a context that real-life typists are required to handle. On the other hand, in terms of interactional authenticity, he considers the task highly authentic due to its high language involvement. However, in my view, since the task is not appropriate to the intended score interpretation and "genuine" to the typists' TLU situation and has almost zero situational authenticity, interactions elicited from this task must also be irrelevant. We cannot discuss the degree of its interactional authenticity because the task that applicants engage with is not pertinent to the major duty of typists in the example. This task is not very useful because it is situationally inauthentic and does not provide any information about the ability expected from typists.

From Bachman's evaluations on these two examples regarding their degrees of authenticity, it appears that Bachman sees *interactional authenticity* as independent from the features of the target language use (TLU) task and its relevance to the assessment task. In

Bachman's view, as long as an assessment task engages with a communicative aspect of language ability (e.g., an interview-type task in the target language), it can be considered interactionally authentic. This suggests that when task developers seek a high degree of interactional authenticity, they should design interactive tasks such as interviews and role-plays, regardless of the context that their tasks should simulate. However, this view may be misleading, indicating that any tasks with low language ability use such as typing from handwritten documents are not desirable in terms of their interactional authenticity, although some may elicit entirely appropriate interactions to their relevant contexts (e.g., the typing task for selecting a proficient typist).

In summary, Widdowson's authenticity holds an interlinked relationship between the context and the processes involved in responding to the context, and from this perspective, interactional authenticity should be affected by the level of situational authenticity. This requires a new interpretation of Bachman's (1991) notion of interactional authenticity to refer to "the extent and type of involvement of test taker's language ability in accomplishing a test task" *that is relevant to the features of the TLU situation* (p. 691). Douglas (2000) also echoes this view:

> it is only by making use of the framework of characteristics to analyze the target language use situation and incorporating relevant characteristics into test tasks [, namely, pursuing situational authenticity] that an authentic interaction between the test taker's language ability and the test tasks can be elicited. (p. 72)

2.2.5. Authenticity in the 21st Century: Validation of the Simulated Target Domain

Despite continuing emphasis in practice, there has been less interest in authenticity per se among language testing professionals and a lack of empirical studies in the past decade. For example, the term 'authenticity' never appears in Bachman and Palmer (2010), a seminal language testing textbook in the 21st century, although in their earlier book, Bachman and Palmer (1996), authenticity was one of the test qualities that determined the usefulness of a test. Perhaps language testers are hesitant to frame their research around the conception of authenticity, for any language test performance is "by its very nature inauthentic, abnormal" (Spolsky, 1985, p. 39), and authenticity may have been downgraded in the eyes of professionals as a nontechnical concept that cannot be researched even if it remains a desired quality in the eyes of test users.

Nevertheless, the spirit of authenticity has long been embedded in other language testing concepts. Spolsky (1985) points out that insufficient authenticity in test materials or methods is a threat to "the *generalizability* of results" (p. 31, emphasis added). This viewpoint is adopted in much recent work, for example, Bachman and Palmer's (2010) argument-based validation framework called an Assessment Use Argument (AUA). One of the warrants for score interpretation in an AUA reflects a need to justify if the interpretation about the ability to be assessed is "**generalizable** to the TLU domain in which the decision is to be made" (p. 159, emphasis in original).

Authenticity, along with some other traditional language testing concepts (e.g., reliability, construct), can be explicitly related to argument-based validation (Chapelle & Lee, 2013). For instance, in the validity argument developed for the TOEFL (Chapelle,

Enright, & Jamieson, 2008), a chain of six inferences[3] connects various claims about how the TOEFL score should be interpreted and used, and two of these inferences, the *extrapolation* inference and the *domain definition* inference, can be considered from the authenticity perspective. The extrapolation inference connects the construct assessed by a test to the quality of linguistic performance in the TLU context, and can be supported by "analyses of the authenticity of the assessment" (Kane, 2012, p. 11).

The domain definition inference, which incorporates Mislevy's ECD perspective, links performances in the TLU domain to the observations of performance on assessment tasks. In order to warrant this connection, the following statement needs to be justified: Observations of performance on assessment tasks reveal relevant knowledge, skills, and ability in situations representative of those in the TLU domain. Domain analysis (e.g., Butler, Eignor, Jones, McNamara, & Suomi, 2000; Rosenfeld, Leung, & Oltman, 2001), expert judgment (e.g., Rosenfeld, Oltman, & Sheppard, 2004), and empirical analysis of task simulation can substantiate the relevance of assessment tasks to TLU tasks. These analyses examine the representativeness of assessment tasks and the relevance of language ability in accomplishing those tasks. In this regard, what is pursued in the domain definition inference is parallel to Bachman's (1991) situational authenticity and the revised definition of interactional authenticity, explained in sub-section 2.2.4., and the domain definition inference also views that the situations highly representative of the target domain are a setting for eliciting the target knowledge, skills and ability.

[3] Each inference in the TOEFL validity argument is given a descriptive name such as "domain definition," "evaluation," "generalization," etc. as in Kane, Crooks, and Cohen (1999).

Table 1 presents the domain definition inference in line with Bachman's (1991) definition of authenticity. The italicized words in the domain definition inference column are equivalent to their respective type of Bachman's (1991) authenticity.

Table 1. The domain definition inference in relation to Bachman's (1991) authenticity

	Domain Definition Inference	Bachman's (1991) Authenticity
Task characteristics	Observations of performance on assessment tasks reveal relevant knowledge, skills, and ability in *situations representative of those in the TLU domain*.	*Situational authenticity*: The perceived relevance of the task characteristics to the features of a specific TLU situation
Language ability	Observations of performance on assessment tasks reveal *relevant knowledge, skills, and ability* in situations representative of those in the TLU domain.	*Interactional authenticity*: A function of the extent and type of involvement of task performers' language ability in accomplishing an assessment task

In this dissertation, situational and interactional authenticity of the integrated listening-speaking task (ILST), one of the integrated task types, and the multimedia-mediated speaking task (MMST) I developed was investigated. The research results provide empirical evidence to support or refute the domain definition inference of a validity argument for an academic English proficiency test that contains the types of integrated tasks featured in this study.

2.3. Academic Spoken English Discourse—The Target Domain

The target domain of my integrated tasks is academic spoken discourse in English-medium universities. In other words, I want to be able to claim that students' performance on the tasks reflects their performance in using academic spoken discourse in English-medium universities. Following the ECD perspective (Mislevy, 2011; Mislevy & Haertel, 2006b;

Mislevy, Steinberg & Almond, 2003), I analyzed this target domain to identify key domain characteristics and reflect them in the development of my tasks. Relevant literature was reviewed for this domain analysis, as done in the previous assessment design projects discussed in Section 2.1.2 (i.e., Mislevy & Haertel, 2006a; Mislevy et al., 2014).

In the current section, I will present what was learned from the domain analysis. Although most research on academic discourse has focused on written registers, researchers have more recently begun to extend their interest to spoken discourse, especially academic lectures. Availability of extensive academic spoken corpora such as the Michigan Corpus of Academic Spoken English (MICASE), then, allowed the expansion of linguistic analyses to academic spoken discourse in general including various academic speech events from colloquia to study groups. The first part of this section will review previous research on analyzing linguistic features of academic lectures, and the second part on examining characteristics of academic spoken discourse in general from a representative sample of academic spoken registers including lectures and other spoken registers.

2.3.1. Academic Lectures

The major communicative function of academic lectures is to teach students about a particular subject, and therefore, they are highly informational as in academic prose. On the other hand, as they occur in the spoken mode, lectures consist of frequent real-time elaboration, a typical characteristic of language in situations where the speaker and listener share time and space dimensions. This interface between an oral and written register in academic lectures creates a unique register (Biber, 1988; 1995; Enikõ, 2000; Swales & Malczewski, 2001).

Academic lectures organize ideas and maintain coherence using various types of discourse markers. Chaudron and Richards (1986) characterized these organizational devices as two global types: micro-markers and macro-markers. Micro-markers are the speaker's signals for segmentation (e.g., *well*, *OK*), time (e.g., *at that time*, *after this*), causality (e.g., *so*, *because*), contrast (e.g., *but*, *on the other hand*), and emphasis (e.g., *of course*, *actually*). Macro-markers state the main ideas of the lecture (e.g., *what I'm going to talk about today is…*) or the major transitions in the lecture (e.g., *another interesting development was…*). Flowerdew and Tauroza (1995) provided empirical evidence that discourse markers aid second language learners' lecture comprehension. Differences in teaching styles and lecture content across academic divisions affected variations in the use of discourse markers (Schleef, 2008).

Among macro-markers, an *aside* (an acceptable escape from surrounding topics) is another important unit to contribute to the overall coherence and consistency of the lecture (Strodt-Lopez, 1991). The three major functions are: 1) to resolve apparent contradictions in semantics and/or pragmatics of the lecture, 2) to highlight contrasts between two academic concepts, and 3) to establish the relevance of the lecture to the real world and clarify the scope of the lecture. Asides emphasize the broader context of the lecture and acknowledge diverse topic structure, and this has practical implications on second language listening instruction and materials development. For example, in an academic listening class, students can be exposed to some examples of asides, instructed on their roles, and prepared to comprehend a global structure of academic lectures. Materials developers can incorporate asides into their materials and provide realistic and authentic texts.

Expanding previous research on macro-markers, several researchers launched a line of research using a lexical phrase approach and differentiated the global and local aspects of macro-markers in academic lectures. Lexical phrases are chunks of language with varying degrees of length, and defined as "conventionalized structures that occur more frequently and have more idiomatically determined meaning than language that is put together each time" (Nattinger, 1986, p. 3). DeCarrico and Nattinger (1988) identified lexical phrases that signal the direction of the lecture (global) and the relationships within it (local), and categorized them based on the specific functions. The authors suggested teaching these chunks and associated functions to aid students in predicting the type of information and interpreting the flow of lectures. Their later work recognized the interactional side of academic lectures, though they play a relatively minor role, as well as the transactional side, and stressed that second language learners should also learn the discourse organizers of the interactional phase for improving comprehension (Nattinger and DeCarrico, 1992). In a later empirical study, Khuwaileh (1999) demonstrated that lexical phrases are crucial to student comprehension of lectures.

More recently, "interactiveness" has also been recognized as an important structural element of academic lectures. Camiciottoli (2004) examined three patterns of interactive discourse:

1) Pronoun + modal/semi-modal + main verb (e.g., *We will/'ll talk about…*),

2) Pronoun + want + infinitive (e.g., *I want to look…*), and

3) Let + pronoun + main verb (e.g., *Let me turn…*).

Camiciottoli demonstrated that these lexical structures with personal pronouns guide students through on-going lectures and positively affect comprehension.

As well as the guiding function examined in Camiciottoli (2004), a first person pronoun, 'we' can also function as the representation of groups as found in the following example, "… *risk is a very colloquial part of the way **we** talk about, things that have to do, with health and illness*" (p. 60). The use of personal pronouns per se, specifically the first person pronoun, 'we,' was investigated in Fortanet (2004). 'We' in academic lectures can denote different referents such as 1) a large group of people including the speaker and other people, 2) the speaker and the audience, and 3) merely the speaker.

In addition to various kinds of textual organizers such as lexical phrases and personal pronouns, phonological signals serve as an additional kind of discourse marker in academic lectures. Thompson (2003) included intonational signals as one of the contributing factors for lecture comprehension, and discovered a positive correlation between text-structuring signals and phonological sections. According to her analysis, topic boundaries in a lecture are noticeably marked by both textual markers and phonological variations in pitch, length of pauses, volume, and speed. The end of a meaning-based section is accompanied by extra low termination and a decrease in volume and/or speed, and the start of the following section begins with an exceptionally high pitch. A lengthy pause is present between the two consecutive sections. These phonological segmentations are coupled with the use of diverse text-structuring signals, and together help students map large-scale organization of the lecture.

2.3.2. Academic Spoken Registers in General

One of the traditional and common spoken events in university classrooms is an academic lecture. Earlier studies on the linguistic analyses of academic spoken discourse

were mainly about lectures and relatively anecdotal due to a small corpus size. However, the introduction of large-scale academic corpora such as the Michigan Corpus of Academic Spoken English (MICASE) and the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) Corpus has provided a more diverse and extensive selection of text, and has allowed generalizable analyses of academic discourse in general, not limited to one specific register[4].

One of the large-scale academic corpora, the MICASE, contains 152 transcripts (approximately 1.8 million words in total) of various types of academic speech events: large lectures, small lectures including seminars and student presentations, other classroom events (i.e., discussion sections and labs), and non-classroom events including colloquia, dissertation defenses, meetings, office hours, and study groups. By looking into this wide range of academic speech samples in the corpus, corpus linguists have gained new insights regarding typical linguistic descriptions across different registers within academic spoken discourse.

Many studies using the MICASE discovered that speakers in academic discourse utilize specific linguistic devices to adopt a stance on the subject matter. One of these, hedging, is a common characteristic of academic spoken discourse. Lindemann and Mauranen (2001) found that a lexical item, *just*, is most often used as a hedge, specifically as a minimizer. *Just* often co-occurs with discourse markers and other hedges, and serves a mitigating function. Hedges soften the authority of the speaker on the ongoing discourse and balance power relations among discourse participants (Mauranen, 2001). In her later work, Mauranen (2004) differentiated hedges as expressions of vagueness (e.g., *kind of*, *sort of*, *or*

[4] Following Biber (1994; 1995) and Conrad and Biber (2001), the term *register* used in this section refers to any language variety defined in situational terms. Registers can be defined at a highly specific level such as *introduction sections in biology research articles* or at a very general level such as *academic writing*.

something, and *or so*) and as mitigators (e.g., *somewhat*, *just*, and *a little bit*). Usage patterns of these two hedging groups differ depending on register types: more monologic genres (i.e., lectures) tend to use hedges as a vagueness indicator, whereas in more dialogic genres, hedges are used more as a mitigator.

Another kind of linguistic device for adopting a stance is an evaluative expression, such as those used in argumentation (e.g., *I think you can't make an argument*). Lexical items such as *argument* are used to express the speaker's evaluative position and organize the ongoing discourse (Mauranen, 2003). Bellés-Fortuño and Campoy-Cubillo (2010) particularly investigated the use of "I feel" in reporting intuitions and ideas, and discovered that this stance marker is used more frequently in highly interactive speech events. Using adjectives (e.g., *good*, *important*, *little*, *standard*) and their intensifiers (e.g., *fairly*, *pretty*, *really*) also contributes to taking a specific attitude towards the topic being discussed (Swales & Burke, 2003).

Another typical linguistic characteristic of academic spoken discourse is its frequent use of targeted expressions such as *any questions*, *an issue perhaps worth mentioning*, *my point is*, and *the thing is*. These targeting expressions help the current speaker structure the ongoing discourse; they also grant a potential for other speakers turning the flow of the discourse, so allocate space to all participating members (Mauranen, 2001). Swales (2001) paid special attention to *point* and *thing* in his corpus analysis, and found that these two discussives generally function as a signal of the importance of the immediate discourse.

Common linguistic expressions used in academic spoken discourse can also be examined on an exploratory basis. By manually searching idioms in the MICASE, Simpson and Mendis (2003) identified 238 idiom types and categorized them based on the six

emerging pragmatic functions: evaluation, description, paraphrase, emphasis, collaboration, and metalanguage. Simpson (2004) compared frequently occurring formulaic expressions in the MICASE with those in other spoken corpora, and discovered expressions more frequent in academic discourse (MICASE), such as *I'm gonna (going to) go*, *you could say*, *(it) turns out*, and *look at it/this*, than in general spoken discourse (comparison corpora). Using a more empirically derived measure of utility, Simpson-Vlach and Ellis (2010) compiled the Academic Formulas List, comparable to the Academic Word List (Coxhead, 2000). This list consists of frequently recurrent formulaic expressions (e.g., *and so forth*, *a kind of*, *and/as you can see*, *it doesn't matter*, *come back to*, *by the way*) specifically across a range of academic genres. The formulas are prioritized considering both frequency and functionality. This array of corpus-based analyses altogether has pedagogical implications for selecting which formulaic expressions to teach.

A more comprehensive linguistic description of academic spoken discourse was given by Douglas Biber and his colleagues in their studies using the T2K-SWAL Corpus, a comparable academic corpus to the MICASE. The T2K-SWAL Corpus is a relatively large (about 2.7 million words) and grammatically annotated corpus across the range of academic spoken and written registers. The spoken part of the corpus consists of 251 texts, approximately 1.7 million words, and ranges from classroom (class sessions and labs/in-class group) to non-classroom registers (office hours and study groups).

Multidimensional (MD) analysis was used in the T2K-SWAL Corpus studies to characterize academic registers. This analytic approach originated in Biber's (1988) work that identified five main dimensions of variation in texts from a compilation of general corpora:

Dimension 1: Involved versus informational production,

Dimension 2: Narrative versus nonnarrative discourse,

Dimension 3: Situation-dependent versus elaborated reference,

Dimension 4: Overt expression of persuasion, and

Dimension 5: Nonimpersonal versus impersonal style.

Using factor analysis, he determined the groups of linguistic features associated with each dimension and developed a statistical technique of MD analysis to discover and interpret the patterns of linguistic variation found in a corpus.

This analytical method led to the detailed description of academic registers in the T2K-SWAL Corpus, conducted by Biber, Conrad, Reppen, Byrd and Helt (2002). In their analysis, the oral university registers consistently differed in their features from the written ones. Academic spoken discourse was characterized by features of 1) involvement and interaction, 2) generally narrative styles, 3) situated reference, 4) more overt persuasion, and 5) fewer features of impersonal nature. When limited to classroom teaching, this register was also found to be involved and interactive, and linguistically similar to face-to-face conversation (Reppen, 2004).

Soon after Biber et al.'s (2002) study, Biber (2003) provided a brief introduction of a new MD analysis of academic language discourse. The inventory of linguistic features was expanded from that of Biber (1988), and this expansion yielded newly identified dimensions specifically for academic discourse: 1) oral versus literate discourse, 2) procedural versus content-focused discourse, 3) narrative orientation, and 4) academic stance. As expected, spoken registers in the university context contained features of more oral than literate language, and also features of more procedural than content-focused language, especially for

classroom sessions in engineering, business, and education. In addition, academic spoken discourse exhibited generally more narrative characteristics, except for classroom teaching in the disciplines of engineering and social science. Finally, features of academic stance were used primarily in the instructor-controlled spoken registers (classroom teaching and office hours).

Biber, et al. (2004) is by far the most comprehensive empirical research on the linguistic characteristics of academic registers. The researchers first analyzed the distribution of several linguistic features such as grammatical categories, lexicogrammatical associations, stance features, vocabulary, and lexical bundle types. MD analyses using both the Biber 1988 framework and the new approach were also conducted to identify the prominent linguistic patterns found in academic registers, and produced the same results as in Biber et al. (2002) and Biber (2003).

Among many language aspects studied in the T2K-SWAL Corpus projects, lexico-grammatical features used for the expression of stance were investigated in Biber (2006). In the context of classroom teaching, modal verbs, especially prediction modals (e.g., *will* and *would*), were the most common grammatical device used as a stance marker. Stance adverbs, specifically for certainty such as *actually* and *in fact*, were also frequently used in classroom teaching. 'Certainty verb + that-clause' (e.g., *I know that* …) and 'desire verb + to-clause' (e.g., *I'd like to* …) are additional expressions of stance, commonly occurring in the classroom teaching context.

Another strand of studies on academic spoken English discourse investigated student and professor opinions on the relevance and importance of academic tasks test developers pre-selected in the successful completion of academic work. For instance, in Rosenfeld,

Leung & Oltman (2001), 370 faculty and 345 students from 21 universities in North America rated 42 task statements developed from the TOEFL theoretical frameworks of four skills. Of 42, four tasks were regarding a speaking skill, and particularly, the task of explaining/informing was recognized as the most relevant to and important for achieving academic success in North American universities. The study finding warranted the use of explaining/informing tasks in the development of test specifications and assessment measures for prospective international students.

Two of the frequently used functions of explaining/informing tasks in academic disciplinary fields are classifications and definitions (Mohan & Slater, 2005; Ogborn, Kress, Martins, & McGillicuddy, 1996). The language associated with these functions consists of "verbs such as "be," both as a relating process and in talking about existence, as in "there are"" (Slater & Butler, 2015, p. 14). The following text from the MICASE showcases the use of the "To be" verb (highlighted in bold and numbered) for classifying and defining academic concepts. The text is from an undergraduate-level second language acquisition class and part of a student presentation of a group research project on Japanese English as a Second Language (ESL) learners' refusals to invitations.

> … after we got the response, we uh tried to make a chart out of, we tried to categorize the responses, and um we made, six, categories, and... and... we tried to um, define the terms... by ourselves, and <PAUSE:19> um... well first you see, (1) **there are** six (not yet,) six categories, and (2) the first one **is** acceptance and, we s- defined it as accepting the offer even though you don't want to, and (3) justification, which **is** citing a specific obligation that precludes accepting the, precludes accepting the invitation an excuse, giving a vague reason, or reasons for not accepting, for example,

> sorry i'm busy, and (4) concession which **is** saying it's not possible for you to accept the invitation, without offering an explanation or reason, or expressing an interest in meeting at some future time. for example i can't make it or, i can't make it but, why don't we meet some other time. and (5) the next one **is** avoidance, um not responding to the invitation or delaying response and (6) the last one **is** refusal, direct rejection of the invitation citing lack of interest for example i don't wanna go … (Simpson, Briggs, Ovens, & Swales, 2002, emphasis added).

The presenter in the text used mainly the "To be" verb including "there are" to explain the six classifications of refusals and the definitions of each. The assessment tasks of the current study were designed to have students orally explain two aspects of an academic concept, and a successful response to the tasks is expected to include frequent occurrences of the "To be" verb used for explanations.

One of the key characteristics of academic lectures is, as discussed in the previous sub-section, interactive discourse using personal pronouns such as "we" and "you" to connect to the audience (Camiciottoli, 2004; Fortanet, 2004). The same feature is also noticeable in the academic spoken discourse of college students. For instance, in the discourse of graduate teaching assistants (TAs), it was found that TAs frequently used personal pronouns when explaining academic concepts (Levis, Levis, & Slater, 2012), and this pronoun use was primarily for interpersonal connections (Slater, Levis, & Levis, 2015). The undergraduate student in the text introduced above also used a personal pronoun, "you," to get the audience's attention and continue the explanation: "… well first **you** see, there are six (not yet,) six categories …" (Simpson, Briggs, Ovens, & Swales, 2002, emphasis added).

To sum up, the current section reviewed previous research on academic spoken English discourse, the target domain of the two integrated assessment tasks, the ILST and the MMST, investigated in this study. The discourse of academic lectures is informational as in the written register, but at the same time, includes frequent real-time elaboration, a typical characteristic of the spoken register. In addition, academic lectures frequently utilize discourse markers such as micro-markers, macro-markers, asides, and lexical phrases, and are known for their interactiveness. The rhetorical division of an academic lecture is marked by phonological signals. These key characteristics were reflected in the input lectures of the ILST and the MMST.

The essential characteristics of academic spoken registers in general are involvement and interactiveness and linguistic similarity to those of face-to-face conversation. The most relevant and important academic spoken task recognized by university professors and students was the task of explaining/informing, and this language use task involves frequent use of the "To be" verb for classifications and definitions and personal pronouns for interpersonal connections. These features were reflected in defining the construct the ILST and the MMST intended to measure and describing the characteristics of the expected response to the assessment tasks.

2.4. Assessment Task Characteristics—Situations Representative of the Target Domain

Knowledge learned from the domain analysis (Section 2.3) was taken into account in task development for the current research. In this research, an assessment task was considered a vehicle for simulating an authentic situation comparable to the target domain, and therefore, during the task design phase, it was important to detail the characteristics of

task that contribute to high situational authenticity. In language testing, Bachman and Palmer's (1996; 2010) task characteristics framework is widely used for describing aspects of tasks, which was also used in the current study.

Bachman and Palmer's framework is comprised of five facets: 1) setting, 2) rubric, 3) input, 4) expected response, and 5) relationship between input and expected response. First, the *setting* is the circumstance under which language use takes place. Its characteristics include physical characteristics (e.g., locations), participants, and time of task. Second, the *rubric* provides the context where tasks are performed and contains the characteristics that give the purpose and structure for tasks and that indicate how language users proceed in accomplishing the tasks. Third, the *input* is the material contained in the task that language users are expected to process and respond to, and fourth, the *expected response* is comprised of the linguistic and/or non-linguistic behavior the task is designed to elicit. Test developers consider the format and language of these two task aspects. Here, characteristics examined from domain analysis play a significant role in describing the linguistic details of task input and expected response. Lastly, the *relationship between input and expected response* concerns interactiveness among participants and/or materials, scope of input that must be processed, and type of information needed for successful task completion.

This task characteristic framework provides a list of task features that developers must consider simulating the target domain to the extent possible, and a system to incorporate the domain knowledge into task development. Aspects featured in the framework can also be used to evaluate how successful an assessment task models the target language use task, namely the degree of situational authenticity.

2.5. Defining Language Ability: An Interactionalist Perspective

Domain analysis informs the description of both task characteristics and the language ability a task requires. When it comes to characterizing a language task, Bachman and Palmer's (1996; 2010) framework provides a straightforward yet relatively sophisticated system of description (refer to Section 2.4). On the other hand, language ability is a highly abstract concept that does not have a readily available tool for its description, but for assessment purposes, it needs to be defined in concrete terms. We call this defined ability a *construct*, which "provides the basis for a given assessment or assessment task and for interpreting scores derived from this task" (Bachman & Palmer, 2010, p. 43).

In the field of measurement, three fundamental theoretical perspectives to construct definition exist: trait, behaviorist, and interactionalist (Chapelle, 1998). A trait perspective defines a construct in terms of the language knowledge and fundamental processes of the test taker, which are expected to remain stable over a range of situations. A behaviorist definition of a construct refers to the features of the context in which a task is performed, limiting the relevance of test performance to a defined context. Finally, an interactionalist definition includes both traits and contextual features, and their interaction mediated by strategic competence of the test taker as shown in Figure 1. An interactionalist perspective seeks the middle ground between the trait and behaviorist perspectives by defining learner capacities in view of a defined range of contexts.

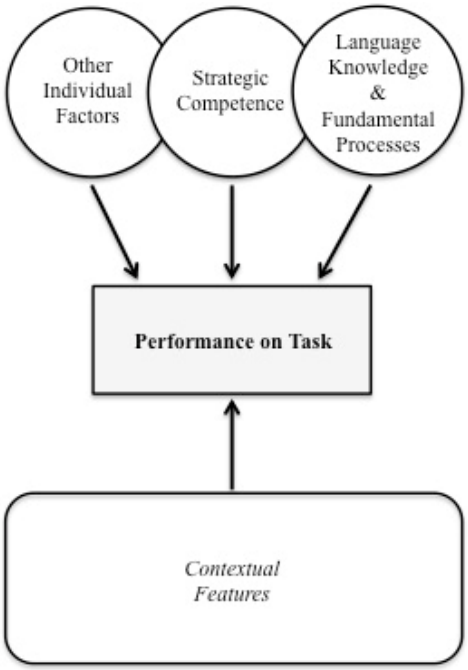


Figure 1. The interactionalist construct definition (adapted from Chapelle, 1998, p. 47)

This approach is consistent with current theory in applied linguistics (Bachman & Cohen, 1998; Bachman & Palmer, 2010), and useful for describing language ability, which is composed of both language knowledge and strategic competence interacting in the context of language use; therefore, in the current study, I used the interactionalist perspective to explicate the construct measured by the two assessment tasks, the integrated listening-speaking task (ILST) and the multimedia-mediated speaking task (MMST).

2.5.1. Language Knowledge, Context, and Systemic Functional Linguistics

From the perspective of the interactionalist construct definition, the context of language use influences the linguistic choices a language user can make during task performance, which can be described by Halliday's (1978) model of language in context. When task performers produce a meaningful sample of linguistic performance (a text) in a given task, they make choices from their knowledge of the language system in order to meet

the demands of the context in which they create the text. Halliday and his colleagues in the school of systemic functional linguistics (SFL) define context theoretically as composed of three parts: *field*, *tenor* and *mode*. Field refers to what is going on in a situation, tenor means who is involved, and mode is what part a text is playing. The language system in Halliday's theory offers resources for constructing the types of meanings that can be realized in view of the contextual configuration, which is defined by the three contextual components. This language system is seen as having meaning-making potential, and an ability to use this potential expands as language proficiency develops (Derewianka, 2001; Mohan & Slater, 2005). In this sense, the ability to use the system is similar to the notion of language knowledge in the interactionalist model of construct definition. Figure 2 presents a visual integration of the interactionalist approach with the Hallidayan model.

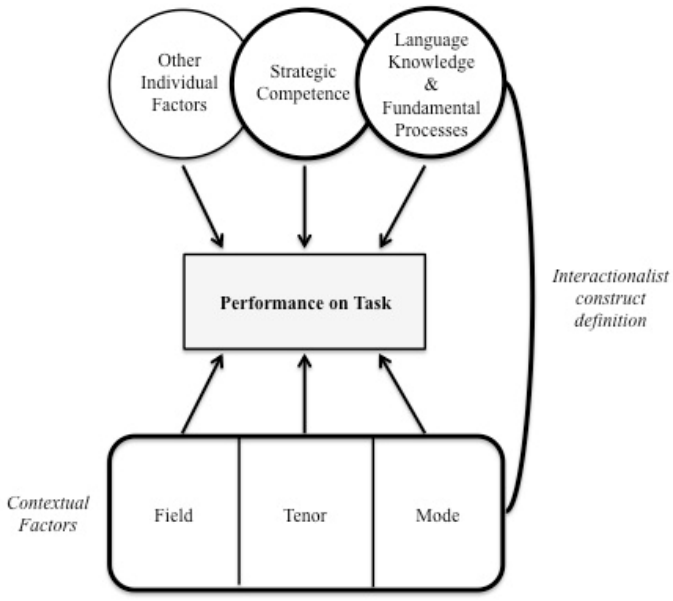


Figure 2. The interactionalist construct definition combined with the Hallidayan model (adapted from Chapelle, 1998, p. 47)

The arch on the right side of the figure connecting language knowledge and the three contextual factors in the figure represents a direct relationship between the linguistic

resources required for performance in a particular context and the context, as defined by the Hallidayan model. In other words, for any construct definition, the contextual features will affect the language knowledge that is needed for successful performance.

2.5.2. Strategic Competence in Language Testing

Strategic competence is another one of the individual factors in an interactionalist construct definition. Canale and Swain viewed the role of strategic competence within the communicative competence model as "to compensate for breakdowns in communication due to performance variables or to insufficient competence" (Canale & Swain, 1980, p. 30) and "to enhance the rhetorical effect of utterances" (Canale, 1983, p. 339). However, Bachman (1990) noted that this facilitative function was limited in defining strategic competence in that it did not describe the operational mechanisms of the concept, and hence he proposed a more general description adapted from Færch and Kasper's (1983) psycholinguistic model of speech production. The model includes three components—assessing what is said, planning utterances, and executing plans. This model treats strategic competence as a cognitive ability for managing communication.

The cognitive approach to conceptualizing strategic competence is refined by embracing a general cognitive theory of human intelligence (Sternberg, 1985, 1988). Strategic competence is defined as a set of "higher-order metacognitive strategies that provide a management function in language use" (Bachman & Palmer, 2010, p. 48), and controls the integration of the individual attributes (e.g., language knowledge) when assessing language use situations. Three general areas in which metacognitive strategies operate are: 1) goal setting (deciding what one is going to do), 2) assessing/appraising (taking

stock of what is needed, what one has to work with, and how well one has done), and 3) planning (deciding how to use what one has) (Bachman & Palmer, 1996; 2010).

Since metacognitive strategies are mental processes, they are not directly observable and therefore need to be inferred by researchers through introspective methods. Therefore, in the current study, I looked for evidence of strategic competence as "*reported* actions and thought processes" from the participants, following the definition used in a recent strategic behavior study in language testing conducted by Swain, Huang, Barkaoui, and Brooks (2009) (p. 2, emphasis in original). Swain et al. (2009) investigated the strategic behaviors 14 graduate and 16 undergraduate Chinese students reported during the speaking section of the TOEFL iBT™. Stimulated recall, an introspective method for reviving participants' memories to recall their thoughts during a certain segment of the immediate past, was used to collect participant verbal reports on their strategy use. The findings suggested that strategy use was integral to performing tasks in the speaking section of the TOEFL iBT™, although the relationship between strategy use and test performance was varied due to complex interactions among characteristics of test takers and tasks. As having worked successfully in Swain et al. (2009), verbal reports on participants' behavior and thinking in the current study should also provide adequate evidence to interpret their goal-oriented cognitive processes during task engagement.

2.6. Visuals in Integrated Assessment Tasks

The discussion in this chapter so far has focused mainly on the theoretical bases used to investigate the two integrated assessment tasks—the integrated listening-speaking task (ILST) and the multimedia-mediated speaking task (MMST)—presented in the current study.

In this section, we shift to considering the format of the tasks themselves, especially examining the use of visuals in integrated assessment tasks. The first sub-section will discuss what role visuals play in defining the construct of integrated assessment tasks, and the second sub-section will explain how the presence of visuals affects task authenticity.

2.6.1. The Role of Visuals in the Construct Definition

A widely used format of current integrated speaking tasks has the task-taker listen to an audio and then speak about its contents (Brown et al., 2005; Frost et al., 2012; Huang, 2010; Inoue, 2009; Iwashita et al., 2008; Lee, 2006; Sawaki et al., 2009). This task format is a substantial advance over separate listening and speaking tests in that it supports a progressive movement in language testing suggesting that "listening and speaking are theoretically and practically very difficult to separate" (Douglas, 1997, p. 25). Although it manages to combine the two skills assessed in a single task, the listen-and-then-speak format still assumes that listening and speaking are two separate constructs that are united in the form of an integrated speaking task, not two components of one construct: *oral communication*. Rather, the listening-to-audio part of this integrated task simply adopts a traditional testing format of listening comprehension and uses an audio recording as a stimulus for the speaking part.

However, this approach implies that the receptive side of oral communication in real life on which the assessment task is modeled is conducted primarily aurally. In assessing listening comprehension, test developers have favored an audio text over a text that includes both oral and visual information for two reasons. First, it requires more resources to develop and administer multimedia assessments (Wagner, 2010b). Second, little research has yet

unraveled the role of visuals in a second language (L2) listening construct (Wagner, 2010a, 2010b).

On the other hand, the use of visuals becomes less disputable (G. Ockey, personal communication, January 27, 2012) and makes more sense if the construct is defined in a manner that makes the interactionalist nature evident: academic speaking in the classroom. A setting where both aural and visual cues are available can promote oral communication more relevant to real life. Since visual information plays a role in many oral communicative events, assessment of learners' abilities to perform in such tasks is needed. In this study, I developed the MMST, which provides visual stimuli along with the aural, to approximate such real-life communication.

According to Bejar, Douglas, Jamieson, Nissan, and Turner (2000) and Ginther (2002), there are two types of visuals: context and content. Context visuals provide information about the setting, the participant(s), and the discourse type, whereas content visuals are directly related to the subject matter of the aural text. The effects of visuals in L2 assessment have been studied mainly in listening comprehension. Some studies show that visuals, mostly in the format of video, are helpful in understanding content (Ginther, 2002; Sueyoshi & Hardison, 2005; Wagner, 2006, 2010b) and preferred by test takers (Progosh, 1996), while other studies find that the test taker's level of comprehension does not significantly differ in testing conditions, either audio texts or videotexts (Coniam, 2001; Gruba, 1993; Londe, 2009).

Methodological shortcomings in some of these studies, for example, incompatible group characteristics, low test reliability, and small sample size (Wagner, 2010b), may have produced the seemingly conflicting findings. However, the function of visuals in L2 listening

may have indeed been a contributing or at least partially contributing factor to the varying findings. When the type of visuals is carefully considered in the studies of visual use in L2 listening comprehension tests, context visuals that do not directly complement audio (e.g., a so-called "talking head") are of less help in comprehension (Coniam, 2001; Gruba, 1993; Londe, 2009). In fact, context and content visuals that help to interpret the meaning of aural texts appear to aid test takers with their understanding (Ginther, 2002; Wagner, 2006, 2010b). One of the studies presenting the ineffectiveness of videos in comprehension, Coniam (2001), acknowledges that "[f]or a videotext to be advantageous, it would appear that the videotexts need to contain more clues than purely paralinguistic ones: the video clues themselves need to be exploitable" (p. 11). That is, the usability of visuals seems to be critical to understanding the content of a video. Visuals that do not directly support what is being said in the video, such as "purely paralinguistic" ones, would not assist in comprehension (Coniam, 2001, p. 11).

Therefore, if we use visuals that facilitate understanding in the stimuli of integrated speaking tasks, it seems that this would improve task-takers' comprehension of information. Visual information can play a crucial role in signaling organization of a lecture and assisting comprehension (Bamford, 2004; Charles & Ventola, 2002; Khuwaileh, 1999; King, 1994). Listen-and-speak tasks, the currently most-widely used integrated speaking task, leaves us uncertain about which skill causes test takers' unsatisfactory performance, either lack of comprehension or production ability (Douglas, 2010), but this muddiness may be alleviated if visuals included in the video material assist the comprehension aspect (G. Ockey, personal communication, January 27, 2012).

2.6.2. The Role of Visuals in Task Authenticity

Adding visuals can also increase authenticity that integrated tasks seek. Multimedia tasks including "rich multimodal input in the form of full motion video," can potentially enhance "authenticity of both input and response" (Chapelle, & Douglas, 2006, p. 9). In the current study, I examined the degree of authenticity of the multimedia task, the MMST, and compared it with that of the ILST, a widely used integrated speaking task, in order to empirically test if the adding of visuals in a stimulus material increases task authenticity.

When the input channel and form of my assessment tasks are compared with those of the TLU tasks that require conveying key ideas of what an individual learned from a lecture, a video lecture in the MMST provides more similar contexts to the TLU tasks (therefore, a higher degree of situational authenticity) than an audio lecture in the ILST does. In real-life contexts, students are rarely given only aural information from lectures unless they intentionally close their eyes during the entire lecture, or they listen to an audio recording of a lecture they missed or want to review for an upcoming test.

My understanding of the relationship between situational and interactional authenticity (refer to Section 2.2.4) suggests that these different degrees of situational authenticity will lead to different extent and types of language ability engaged in the assigned task, namely, different degrees of interactional authenticity. As the contents of the spoken responses rely extensively on those of the input lecture, changes in the delivery mode would be expected to affect the relationship between the input and the expected response. Visual information of both context and contents available in the MMST input should encourage the MMST performers to pay attention to the relevant key ideas of the input lecture, and incorporate this recognized information into using their language knowledge to produce

spoken responses. The presence of visual information in the MMST input is also predicted to facilitate the use of visuals to understand, think about, or remember information, and affect the task takers' strategy of storing, appraising, and selecting resources needed to complete the assigned task. In cognitive psychology, it has been known that a visually presented text can be transformed into the phonological store of working memory (Gathercole & Baddeley, 1993), and information obtained through the visual channel is more likely to be remembered than information obtained from auditory input (Miller & Burton, 1994), known as pictorial superiority effect (Levie, 1987). From this perspective, the ILST performers, influenced by their aural only input, are expected to be less involved in using the language ability defined in the construct than the MMST performers.

2.7. Construct Definition of the ILST and the MMST

The construct intended to be measured by both the ILST and the MMST is defined as an interactionalist integrated speaking-listening construct. The goal of these assessment tasks is to measure the test taker's ability to appropriately and intelligibly convey key ideas from a lecture segment representative of academic course content in response to questions. This explaining/informing function is highly important for achieving academic success in North American universities (Rosenfeld, Leung & Oltman, 2001), and therefore their appropriate measurement needs to be better understood. In the interactionalist approach to construct definition, contextual features of tasks affect the language ability of the task performers as described below.

2.7.1. Contextual Features Simulated in the ILST and the MMST

In the interactionalist approach, the context dimension is comprised of three theoretical components: *field* (what is going on in a situation), *tenor* (who is involved), and *mode* (what part a text is playing). *Field* of the ILST and the MMST involves the social process of explaining academic subjects. The interactants engaged in this activity are college students and a professor, whose introductory-level lecture they listen to and then which they are called upon to answer questions about. The rhetorical/pragmatic function of the lecture is to classify and define some related concepts of an academic but not too technical topic. After (or sometimes during) the lecture, the students use points and examples from the lecture and talk about what they understood as they might do in an academic situation for a comprehension check or as part of further discussion with their classmates and/or professor. Lecture topics determine what the students explain, and in the current study, two different topics, one from biology and another from business administration, create different contextual configurations in terms of *field* for both the ILST and the MMST.

The only difference in *field* between the ILST and the MMST is the materials given in the lecture. In the ILST, the professor is portrayed in a still picture, and the lecture is delivered only aurally. On the other hand, in the MMST, a video-taped professor gives a lecture using supplementary PowerPoint slides, and the students gain both aural and visual information in understanding the lecture.

The *tenor* in the interactionalist construct definition means the role relationships established by the interactants in a given context. The students in the ILST and the MMST do not have an intimate relationship with the professor in the lecture. They are expected to take a neutral stance on the lecture and adopt their speech role as an information provider to the

audience (imaginary classmates and/or the professor) in an institutional setting, where the interlocutors are communicating in order to teach and learn.

The *mode* of the ILST and the MMST is characterized by the following elements: as the language used in the context is oriented towards expounding some domain of academic knowledge, the rhetorical mode is describing; the lecture and student explanations occur in sequence as if the interaction were an asynchronous dialogue consisting of the professor lecturing to the students who, in turn, speak to the (imaginary) audience.

However, the medium and channel of the lecture mode in the ILST and the MMST differ. As mentioned earlier, the mode of the ILST lecture is aural (spoken and phonic), whereas the MMST lecture consists of diverse semiotic systems such as spoken, written (e.g., key phrases from the lecture), and visual (e.g., figures of examples used in the lecture). While the role of language in the lecture differs between the ILST and the MMST, language is the only means in students' oral explanation for both assessment tasks. Table 2 below summarizes contextual features the ILST and the MMST intend to simulate.

Table 2. Contextual features simulated in the ILST and the MMST

Contextual Factors	Contextual Features	
Field	• The social process of explaining academic subjects	
	• The rhetorical/pragmatic function of classifying and defining some related concepts of an academic, but not too technical topic	
	• (ILST) The aural lecture materials	• (MMST) The aural and visual lecture materials
Tenor	• College students and a professor	
	• A neutral stance on the lecture	
	• A speech role as an information provider to the audience in an institutional setting	
Mode	• An introductory-level lecture and student explanations in sequence	
	• Language as the only means in students' oral explanation	
	• (ILST) The spoken and phonic medium and channel of the lecture	• (MMST) The spoken, written and visual medium and channel of the lecture

The differences between the ILST and the MMST are the field and mode of the input lecture—the aural lecture for the ILST in contrast to the aural and visual lecture for the MMST.

2.7.2. Language Ability Elicited by the ILST and the MMST

The language ability assessed on the ILST and the MMST is composed of language knowledge and strategic competence in an academic context (Bachman & Palmer, 2010). The academic context is sampled and defined in the assessment tasks to elicit certain aspects of language ability.

The important characteristics of the academic tasks are defined by the three contextual variables, *field*, *tenor*, and *mode*. *Field* of the ILST and the MMST determines linguistic features the students are asked to comprehend and use. Successful task performers are expected to adequately understand the main ideas and details of the lecture and use relevant information in their oral explanation of the concepts discussed in the lecture. Their key vocabulary should be closely related to the topic and examples of the lecture. Since successful task completion requires classifying and defining different concepts, a competent performer would have frequent syntactic structures such as “there are A and B” for classification and “A is B” and “A means B” for definitions.

The *tenor* creates a context of the ILST and the MMST where task performers define their role as an information giver, sharing no personal opinions. This encourages successful task performers to use primarily statements and less attitudinal language such as modal finites (e.g., must) and mood adjuncts (e.g., definitely).

The *mode* of the ILST and the MMST influences the organization of oral explanations offered by task performers. The ILST and the MMST performers are expected to construct their responses addressed to the (imaginary) audience. A competent performer would frequently use pronouns as exophoric reference (e.g., "we/our/us" referring to the speaker and the imagined audience, "you/your/you" referring to the imagined audience, and "I" referring to the speaker) for addressing the audience to their explanations. Due to additional semiotic systems used in the MMST lecture such as supplementary PowerPoint slides, those who perform the MMST have more assistance in comprehending the lecture and thus, presumably, would be better able to organize key ideas in their responses.

In terms of strategic competence, all three areas of Bachman and Palmer's (2010) strategy use, 1) goal setting, 2) assessing/appraising, and 3) planning, are used for successful task completion. Task performers seek and identify task goals based on their knowledge of what happens in a classroom, and do purposeful comprehending and talking about the lecture. While receiving information from the lecture, successful task performers are expected to use strategies to anticipate the content according to their comprehension of the lecture and/or common knowledge. They also use notes to remember and/or organize information, and imagery, either generated or actual, to understand, think, or remember information for successful task completion, and make good use of these in planning their oral explanations. It is expected that the MMST performers would frequently use the strategy of using imagery, as visual information in the MMST input lecture would affect task performers in comprehending the content and storing, appraising, and selecting resources needed to complete the assigned task. Finally, successful task performers evaluate if their comprehension and language production were successful. Table 3 below summarizes

language ability, consisting of language knowledge and strategic competence, the ILST and the MMST intend to elicit.

Table 3. Language ability elicited by the ILST and the MMST

Language Ability Components	Language Ability Elicited
Language knowledge affected by	
Field	• Adequate understanding of the main ideas and details of the lecture • Use of relevant information in the oral explanation of the concepts discussed in the lecture • Key vocabulary closely related to the topic and examples of the lecture • Frequent syntactic structures such as "there are A and B" for classification and "A is B" and "A means B" for definitions
Tenor	• Primary use of statements • Less use of attitudinal language such as modal finites (e.g., must) and mood adjuncts (e.g., definitely)
Mode	• Use of pronouns as exophoric reference • Organization of key ideas in oral explanation
Strategic competence	
Goal setting	• Seeking and identifying task goals • Doing purposeful comprehending and talking about the lecture
Assessing/appraising	• Anticipating the content • Using notes to remember and/or organize information • Using imagery, either generated or actual, to understand, think, or remember information • Evaluating if comprehension and production were successful
Planning	Making good use of what are assessed/appraised

Although both the ILST and the MMST intend to elicit the same aspects of language ability summarized in the table, due to some differences in contextual features the two assessment tasks simulate (e.g., the medium and channel of the input lecture), the extent of elicited language ability for certain aspects may vary between the two tasks.

2.8. Research Questions

The purpose of this dissertation is to evaluate both the *situational* and *interactional authenticity* of the two innovative assessment tasks for academic English proficiency, the integrated listening-speaking task (ILST) and the multimedia-mediated speaking task (MMST). The ILST is currently a widely used integrated speaking task, whereas the MMST has been developed specifically for this dissertation research with the intention of increased authenticity of response as well as input. The degree of the two types of authenticity between the ILST and the MMST was compared.

As another way to analyze the interactional authenticity, this dissertation also examined the construct of language ability used to perform on the ILST and the MMST. Language ability involved in the ILST and the MMST was described from the interactionalist perspective of construct definition (Chapelle, 1998). The detailed construct analysis is indispensable for investigating interactional authenticity due to the complex construct these innovative assessment tasks measure (Frost, Elder, & Wigglesworth, 2012; Plakans, 2013; Suvorov & Hegelheimer, 2013; Yu, 2013). The elicited language ability of the ILST and the MMST was compared.

Considering these goals, the study address the following research and sub-research questions:

1. **To what degree do examinees, who are students in the TLU context, perceive each of the two assessment tasks—the ILST and the MMST—as corresponding to the target language use (TLU) task, explaining/informing tasks in the academic domain? How different is this perceived situational authenticity between the ILST and the MMST?**

The first research question asks students' perception of *situational authenticity* regarding correspondence of task characteristics between the ILST/MMST and the target language use (TLU) task—explaining/informing tasks in the academic context. **A**ny difference of perceived situational authenticity between the two task types, the ILST and the MMST, was considered.

2. **To what degree do examinees consider the involvement of language ability, defined in the construct, to be high in accomplishing each of the two assessment tasks? How different is this perceived interactional authenticity between the ILST and the MMST?**

The second research question asks the examinees' perception of *interactional authenticity* regarding the involvement of language ability in accomplishing the two assessment tasks. Any difference of perceived interactional authenticity between the two task types was considered.

3. **What language knowledge do the two assessment tasks—the ILST and the MMST—elicit in the examinees' spoken response? How different is this elicited language knowledge between the ILST and the MMST and across assigned task scores?**

The third research question sought to empirically analyze what language knowledge the ILST and the MMST measure, in view of the interactionalist perspective on defining a construct (Chapelle, 1998). The elicited language knowledge was compared between the ILST and the MMST and also across assigned task scores.

The third research question, particularly, includes three sub-questions examining the relevant areas of language knowledge defined in the task construct.

3a) How different is the use of processes between the two assessment tasks? Are there differences affected by task scores?

The first sub-question concerns the use of *processes*, the activity or way of being in an event (e.g., *use*, *expect*), between the task types. Any remarkable differences in *process* use across assigned task scores were also considered to determine if this language knowledge develops as proficiency progresses.

3b) How different is the use of processes between the task versions, biology vs. business administration?

This sub-question relates to *Field* (what is going on in a situation), one of the theoretical contextual features in the construct definition. The question investigates the use of *processes* between the two task versions, biology and business administration. A *process* is a set of linguistic resources a task performer can choose from to express ideational meanings to represent *Field*. By defining the *field* as specific topics within biology and business administration, the two input lecture topics are expected to affect examinees' spoken responses in terms of linguistic choices they make to embody the "what" of a situation. This necessitates a comparative analysis between task versions.

3c) How different is the use of exophoric reference between the two assessment tasks? Are there any notable differences across assigned task scores?

This sub-question is associated with *Mode* (what part a text is playing), another one of the theoretical components in the contextual part of the interactionalist construct. In particular, the use of exophoric reference (referring to something outside of the text itself) between the two assessment tasks was examined to check the examinees' awareness of the imaginary audience beyond the immediate context.

4. **How is strategic competence involved in completing the two types of assessment tasks? Are there any notable differences between the two task types? Do these differences vary across assigned task scores?**

The fourth research question sought to empirically analyze what strategic competence the ILST and the MMST measure, in view of using the interactionalist perspective on defining a construct (Chapelle, 1998). The elicited strategic competence was compared between the ILST and the MMST and also across assigned task scores.

2.9. Chapter Summary

This chapter consisted of eight main sections. The first section discussed an approach to test development called evidence-centered design (ECD), specifically its first step, domain analysis. Since an attempt to model the target domain in test development and validation is conceptually equivalent to seeking to construct authentic tasks, the second section provided an overview of authenticity in language learning and testing, showing how the conception of this notion has evolved, and presented the authenticity framework for this dissertation. For understanding the characteristics of the target domain of tasks featured in the current study, the third section reviewed previous studies on academic spoken English discourse. The fourth section explained the task characteristic framework the current study used to describe the context the assessment tasks intended to simulate. The fifth section dealt with a conceptual discussion of the interactionalist perspective on defining language ability. This perspective relates to the construct definition of the two assessment tasks used in this dissertation, the ILST and the MMST. The sixth section demonstrated grounds for using visuals in integrated assessment tasks, which aimed to justify the use of visual information in

the integrated task I developed, the MMST. The seventh section described the construct of the ILST and the MMST, reflecting the theoretical underpinning discussed in the previous sections. The last section of the chapter articulated research questions the current study examined.

CHAPTER 3. METHODOLOGY

In this chapter, the methodological details of the current study will be presented. First, I will explain the two-part research design employed for the study. Second, the participant recruiting procedures and the demographic information of the recruited participants will be described, followed by the details of the materials and instruments used in the study. Next, I will present the procedures of data collection and the details of the data, and explain how the data were analyzed to answer the research questions. Lastly, types and sources of data and analytical methods used to answer each of the research questions will be summarized.

3.1. Two-Part Research Design for Authenticity Analysis

The current study was designed to answer the four research questions introduced in the previous section. The first two investigated the test takers' views on the authenticity of the context created by the two assessment tasks, the ILST and the MMST, and the extent of their language ability involved in accomplishing the tasks. The last two questions examined how the test takers' language ability, defined in the construct definition (refer to Section 2.7), was actually involved in completing the two assessment tasks. These two major research foci demanded a two-part research design, and each part used a different research design, respectively: 1) an embedded design for the first part and 2) a qualitative research design with data transformation for the second. The embedded data model was chosen to investigate how participants perceived the authenticity of 1) the contexts realized by the assessment tasks and 2) language ability involved in each task context. The qualitative research design was selected to examine the manifestation of language ability involved in participant task

performance. Together, the two research designs were employed to evaluate the degree of authenticity of the two assessment tasks. Each of the research designs will be discussed in detail in the following sub-sections.

3.1.1. Mixed-Methods Design: The Embedded Model

Following the methodology of Liu's (2005) comprehensive authenticity investigation, the analysis of perceived authenticity in the current study employed the embedded model of mixed methods research (Creswell & Plano Clark, 2007). This mixed methods design includes two data sources: 1) quantitative data from authenticity questionnaires, and 2) qualitative data from follow-up interviews which elaborate the main quantitative findings. In this study, the quantitative data consisted of ratings on a six-point Likert scale regarding the perceived authenticity of involved task features and language ability. The qualitative data were oral explanations of the choices of ratings assigned to the questionnaire items. Figure 3 shows how the two data sources in this research design were used for the perceived authenticity analysis.

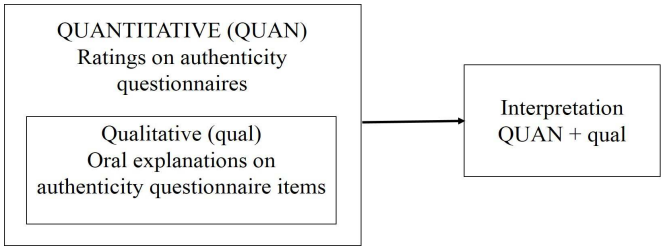


Figure 3. The embedded model used in the study of perceived authenticity

The use of all capital letters of the word, "QUANTITATIVE," in the figure signifies the relative importance of these findings. Qualitative findings embellished the quantitative

questionnaire findings by providing additional details that helped to explain the quantitative results, but the primary focus was on the quantitative results.

3.1.2. Qualitative Research Design with Data Transformation

The third and fourth research questions examined the target construct, framed by the interactionalist perspective (Chapelle, 1998), that the two assessment tasks, the ILST and the MMST elicited, and employed a qualitative research design with data transformation, as done in many construct analysis studies of integrated assessment tasks (e.g., Frost et al., 2012; Gebril & Plakans, 2013; Jin & Mak, 2013; Plakans, 2009; Swain et al., 2009). The elicited target language ability, consisting of language knowledge and strategic competence, can be thoroughly investigated by qualitative data such as speech samples and oral reports of strategy use, and therefore the current study adopted a qualitative research design for the analysis of elicited language ability. Research questions three and four also addressed comparisons of elicited language ability between the two task types, and this comparative nature of the study necessitated transforming qualitative linguistic analysis into quantitative data and investigating the statistical significance of any differences found between the groups.

To answer the third research question about the elicited language knowledge manifested in examinee spoken responses, the transcripts of the speech samples were coded on a number of systemic functional linguistic (SFL) features (Halliday & Matthiessen, 2014) (refer to Section 3.6.2. for details). The SFL framework was used for an analysis that links the contextual configuration to specific aspects of the grammatical system. The number of expected grammatical features in each of the spoken responses was used to interpret what

language knowledge was elicited. Figure 4 visually presents the research design of this language knowledge study.

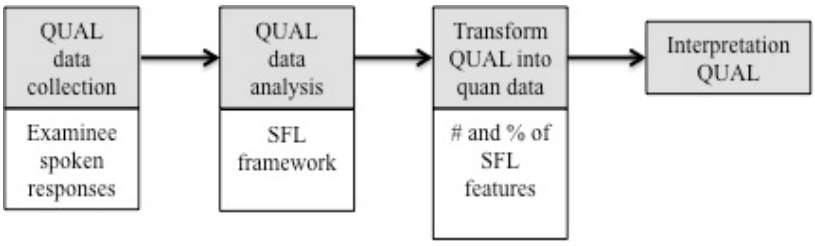


Figure 4. The data transformation qualitative design used in the language knowledge study

The fourth research question also utilized the qualitative research design with data transformation (Figure 5). I obtained examinees' strategic behavior reports from stimulated recall. This introspective method is useful for obtaining evidence of cognitive processes without interrupting participants' task performance (Gass & Mackey, 2000), and in this dissertation, this method provided detailed data which explained how participants tapped their strategic competence to accomplish the two assessment tasks. These qualitative data were transformed into quantitative data by coding the categories of strategies observed in the test takers' oral reports, using a modified version of Swain et al.'s (2009) coding scheme (refer to Section 3.6.4. for details), and the number of strategy categories served as a basis for interpreting the elicited strategic competence, as shown in Figure 5.

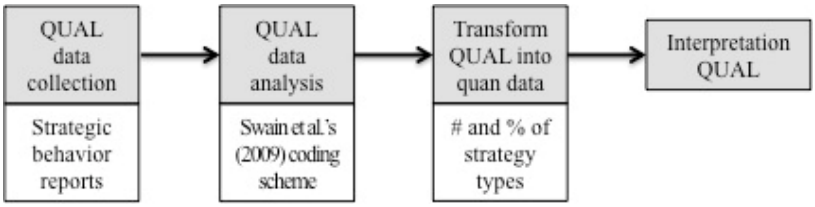


Figure 5. The data transformation qualitative design used in the strategic competence study

3.2. Participants

Participants for this study were 93 international undergraduate and graduate students who were 18 or older and enrolled at a Midwestern research university in the United States (U.S.). Twenty participants were recruited in Spring 2013 and the remaining 73 in Fall 2014. Participants came from four different groups in the university: 1) oral communication classes in an intensive English program (n = 17), 2) English as a Second Language (ESL) listening courses (n = 16), 3) ESL speaking courses for international teaching assistants (n = 33), and 4) a graduate program in applied linguistics (n = 27). The aim of various recruiting sources was to include a range of academic English proficiency levels representative of the population of international students whose first language is not English, ranging from those who were conditionally admitted to the university with a TOEFL total score of 55 or higher (Group 1) to those who were admitted to a university with a TOEFL admissions requirement total score of 111 (Group 4). This was the relevant sample to assess the quality of the tasks for investigating the varying degrees of linguistic knowledge and strategic competence for task performance that represent those for the population.

For Groups 1, 2, and 3, I asked the program directors and course instructors for permission to visit their classes, and during my visits, I asked students in those sections to consider voluntary participation in the study. The exception was that two sections of Group 3 were from classes I taught in Spring 2013 and Fall 2014, respectively. I announced the study to these students during class and recruited research participants, emphasizing that participation in the study was completely voluntary. Group 4 received an email, and only those who agreed participated in the study.

All participants signed an informed consent to indicate their agreement in voluntary participation. Half of the participants in each group were randomly assigned to the integrated listening-speaking task (ILST) and the other half to the multimedia-mediated speaking task (MMST) (Refer to Table 4).

Table 4. Assignment of participants in each task (*n = 93*)

Recruiting Group	ILST			MMST		
	Spring 2013 (*n =10*)	Fall 2014 (*n =36*)	Total (*n =46*)	Spring 2013 (*n =10*)	Fall 2014 (*n =37*)	Total (*n =47*)
Group 1	2	6	8	2	7	9
Group 2	3	4	7	3	6	9
Group 3	2	14	16	3	14	17
Group 4	3	12	15	2	10	12

The ILST group consisted of slightly more females (19 males and 27 females), and the MMST group consisted of slightly more males (26 males and 21 females), showing the gender proportion was relatively balanced. The average age was 26.0 for both the ILST group (*SD* = 5.10; *range* = 18 - 38) and the MMST group (*SD* = 5.61; *range* = 18 - 43). The participants were from diverse academic disciplines (refer to Table 5). The largest proportions came from the following three disciplines: 1) social science, 2) natural science, and 3) engineering. Seventeen participants in each group were students in social science, and eight participants in the ILST group and seven students in the MMST came from departments in natural science such as biology, chemistry, and physics. Seven participants in each group were engineering students, majoring in aerospace engineering, chemical engineering, civil engineering, computer engineering, electrical engineering, or mechanical engineering.

Table 5. Academic disciplines of the participants (*n = 93*)

Academic Disciplines	ILST (*n = 46*)	MMST (*n = 47*)
Agriculture	4	2
Architecture and design	1	1
Business	5	5
Education	1	1
Engineering	7	7
Formal science[1]	2	4
Humanities	1	0
Journalism, media studies and communication	0	1
Natural science	8	7
Social science[2]	17	17
Undecided	0	2

[1] Mathematics, computer sciences, and statistics
[2] The majority came from applied linguistics: 15 of the 17 participants in the ILST group and 12 of the 17 in the MMST group

The average length of residence in the U.S. for the ILST group was 22.9 months (*SD* = 24.46; *range* = 3 – 84), compared to 18.1 months for the MMST group (*SD* = 22.01; *range* = 2 – 108). Chinese was the most frequently spoken first language in each group (22 for the ILST group and 20 for the MMST group), followed by Korean (8 for the ILST group and 10 for the MMST group). Participants of these two first languages were fairly evenly distributed. Overall, the demographic information for each of the conditions was roughly alike.

3.3. Materials and Instruments

In this section, materials and instruments used to elicit data from participants are described. The first two sub-sections explain the details of the two assessment tasks, the MMST and the ILST. Next, the protocol of stimulated recall is explained. The last two sub-sections introduce the two authenticity questionnaires, the situational authenticity questionnaire and the interactional authenticity questionnaire, respectively.

3.3.1. Multimedia-Mediated Speaking Task (MMST)

The multimedia-mediated speaking task (MMST), one of the integrated speaking task types, required participants to watch a short simulated video lecture on an academic topic (approximately two minutes in duration) and then orally explain two main points presented in the lecture. Two versions of the MMST were developed for this dissertation in an attempt to vary topics to examine any influence of topical difference in task performance, one from natural science and the other from the discipline of business. Each task version maintained the same format, but differed in topics, with task version A, biology, and task version B, business marketing. The script of the biology lecture (340 words) was from the TOEFL iBT® Test Sample Questions (Educational Testing Service, 2010) and the business lecture script (307 words) was from the TOEFL iBT® Quick Prep: Volume 3 (Educational Testing Service, 2011)[5]. Participants were not exposed to these lectures prior to participating in this study.

The simulated video lecture involved one speaker (a female in task version A and a male in task version B), who acted as a professor. The speakers for both were native speakers of English. The female was a senior lecturer in the English department, and the male was a postdoctoral researcher in the same department. While being videotaped, the speakers were prompted with the lecture script and naturally said it as if they had been lecturing on these materials in a real classroom with appropriate facial expressions and gestures. Additionally, three types of content visuals (e.g., figures) in PowerPoint format were used to elaborate the lecture:

---

[5] Copyright © 2013 Educational Testing Service. www.ets.org

The TOEFL iBT Test Sample Questions and Quick Prep materials are reprinted by permission of Educational Testing Service, the copyright owner. All other information contained within this document is provided by the author. No endorsement of any kind by Educational Testing Service should be inferred.

1. Illustrating the oral stimulus (e.g., a picture of tools in the biology lecture on two definitions of a tool used by animals);
2. Organizing information in the stimulus (e.g., an outline of the two main points (i.e., the definitions of a tool) of the biology lecture);
3. Replicating the oral stimulus (e.g., a phrase of the two definitions of a tool explained in the biology lecture) (Bejar et al., 2000).

A fourth type of visual, one used to "supplement the oral stimulus," was purposefully omitted since providing information beyond the central focus of the lecture could presumably make comprehension more difficult (Bejar et al., 2000). The two video lectures used six PowerPoint slides for both the ILST and the MMST in the same order of content visual types (Appendix A). The PowerPoint slides were projected to a screen located next to the speakers (refer to Figure 6), and advanced to coordinate with the lecture. When a slide contained text, animation effects were used to show appropriate text one piece at a time and to synchronize them with the corresponding parts of the lectures.

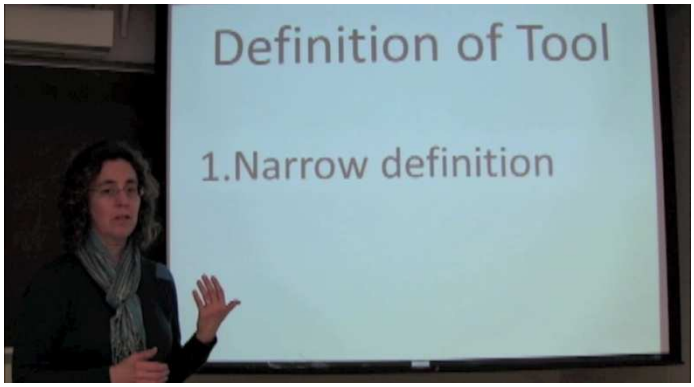


Figure 6. A screenshot of the simulated biology lecture in the MMST

The speakers stayed in one position in the frame while giving their lectures. Their upper bodies were shown to clearly capture their facial expressions and hand gestures. The

video lectures were recorded by using a full high-definition camcorder (Canon Vixia HF M50). A shotgun condenser microphone (RØDE VideoMic) was mounted on the camcorder to obtain high sound quality. Background noise was further removed in Audacity, an audio editing computer software application, and edited in Adobe Premiere Pro CS6, video editing software.

The prototype test tasks were created by embedding these edited video lectures in a set of the PowerPoint slides containing pictures and written instructions synchronized with the audio. Appendix B provides a detailed outline of the slides. While completing the tasks, the participants watched these narrated slide shows on a laptop computer screen with a computer speaker. The slides were set for automatic transitions with the correct pre-determined time intervals in between. The goal was to mimic a computerized test to the extent possible so that participants did not have to manually advance the slides. Appendix C presents the task specifications for the MMST.

The rubrics used to score examinee task performance was the *TOEFL® iBT Test Integrated Speaking Rubrics* (ETS, 2008). The MMST was based on the TOEFL integrated speaking task; therefore these rubrics were used in the current study. The rubrics consist of four holistic score bands from a score of one to four. Each score band is described on the following four aspects: 1) general description, 2) delivery, 3) language use, and 4) topic development.

3.3.2. Integrated Listening-Speaking Task (ILST)

The integrated listening-speaking task (ILST), the other integrated speaking task type, required participants to listen to a simulated short audio lecture on an academic topic

(approximately two minutes in duration) and then orally explain two main points presented in the lecture. The two versions of the MMST video lecture were converted to MP3 files to produce the audio lectures for the ILST. As in the MMST, each task version maintained the same format, but differed in topics, with task version A, biology, and task version B, business marketing.

Following the same development procedures used for the MMST, the audio lectures were embedded in a set of PowerPoint slides containing pictures and written instructions synchronized with the audio. Appendix D provides a detailed outline of the slides. A picture of a talking professor, who was the actual speaker in the audio, was presented while participants listened to the lecture (Figure 7). This provided contextual information, but no content-related details.

Figure 7. A picture of the speaker delivering the biology lecture in the ILST

As in the MMST, the narrated slide shows were played on a laptop computer screen with a speaker. Appendix E presents the task specifications for the ILST. The same rubrics, the *TOEFL® iBT Test Integrated Speaking Rubrics* (ETS, 2008), was used to score examinee task performance on the ILST because the ILST was also based on the TOEFL integrated speaking task.

3.3.3. Stimulated Recall Protocol

Stimulated recall, an introspective method, was used to collect evidence of strategic competence, operationalized as reported actions and thought processes from the participants. The stimulated recall session captured participants' reports by having participants watch the video of themselves doing either the ILST or the MMST and using it as a stimulus to recollect and report what they were thinking at the time they were doing the assessment task.

I designed the following protocol in accordance with the guidelines of Gass and Mackey (2000):

1. I provided procedural instructions to participants including when and how to pause the video before the session began.
2. When participants stopped the video, I listened to what they said. When I stopped the video, I asked general questions such as “what were you thinking here/ at that point/ right then?” When participants responded that they did not remember, I did not ask further questions because answers that are not immediately provided are highly likely based on what participants think at the moment of the stimulated recall session, not during the task, or some other memory or perception.
3. I did not focus or direct participant responses other than providing “what were you thinking then,” and did not react to the responses other than providing backchanelling cues such as “oh,” “mhm,” “great,” “good,” “I see,” “uh-hum,” and “ok,” or no responses.

Appendix F explains the detailed stimulated recall protocol including the role of the researcher and step-by-step instructions.

3.3.4. Situational Authenticity Questionnaire

A questionnaire to examine the situational authenticity of the ILST and the MMST (see Appendix G) was developed by adapting the interpretation questionnaire from Liu's (2005) comprehensive authenticity study. The ten questionnaire items correspond to the aspects of language tasks in the Bachman and Palmer (2010) framework, explained in Section 2.4. The first item concerns the authenticity of task instructions in the rubric. Items 2 and 3 are related to the characteristics of the task setting. The remaining items are input and response-related. Item 9 particularly asks the relationship between the input lecture and examinee spoken responses to gather data on the perceived authenticity of the unique characteristics of integrated assessment tasks. Participants were asked to rate the statements in the questionnaire on a six-point Likert scale regarding the degree of similarity between the ILST/MMST and the TLU task (1 = very different; 6 = very close), so a higher score would indicate that the characteristics of the assessment task are perceived as authentic relative to the features of corresponding TLU tasks, in other words that they are perceived to have a high degree of situational authenticity.

3.3.5. Interactional Authenticity Questionnaire

In addition to the questionnaire for investigating situational authenticity, I developed a questionnaire to examine the interactional authenticity of the ILST and the MMST (see Appendix H). The construct analysis from my pilot study guided the selection of questionnaire items. The 14 items included in the questionnaire correspond to the aspects of the target language ability as defined in the task construct (refer to Section 2.7). The first six items concern the construct-relevant language knowledge as analyzed through the systemic

functional linguistic framework. The remaining items are related to the construct-relevant strategic competence. Participants were asked to rate the questionnaire items on a six-point Likert scale regarding the extent to which each aspect of language knowledge and strategic competence was involved in accomplishing their assigned task (1 = not at all; 6 = a lot), so a higher score would indicate that the linguistic and strategic component of language ability as defined in the construct was greatly involved in accomplishing the assessment task; in other words, that it displayed a high degree of interactional authenticity.

3.4. Data Collection Procedures

The study took place in a quiet conference room on campus. Only one person participated at a time. The participants from each of the four different groups performed one of the four tasks as presented in Table 6. Participants within each of the four groups were randomly assigned to each of the assessment tasks (stratified-random assignment) and relatively evenly distributed.

Table 6. Overview of task assignment distribution

Task assignment	# of assigned participants												
	Spring 2013 (*n* = *20*)				Fall 2014 (*n* = *73*)				Total (*n* = *93*)				
	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	Total
ILST version A, biology	1	2	1	1	2	2	7	7	3	4	8	8	23
ILST version B, business	1	1	1	2	4	2	7	5	5	3	8	7	23
ILST total	2	3	2	3	6	4	14	12	8	7	16	15	46
MMST version A, biology	1	2	1	1	3	3	7	5	4	5	8	6	23
MMST version B, business	1	1	2	1	4	3	7	5	5	4	9	6	24
MMST total	2	3	3	2	7	6	14	10	9	9	17	12	47

Note. G1 = oral communication classes in an intensive English program; G2 = ESL listening courses; G3 = ESL speaking courses for international teaching assistants; G4 = a graduate program in applied linguistics

3.4.1. Task Performance

Each participant carried out one task, either the ILST or the MMST, of either task version A (biology) or B (business) (Refer to Sections 3.3.1 and 3.3.2). Participants were permitted to take notes while they either listened to or watched the lecture, and then were allowed 20 seconds of planning time before producing their oral response within a one-minute time limit. The responses were recorded in Audacity, an audio recording computer software program, connected to a directional USB digital microphone (Telex M-560) for the analysis of speech samples. The whole task completion procedure lasted approximately five minutes for each participant.

For the subset of 20 participants who participated in the study in the Spring 2013 administration, the entire five-minute process of task completion was videotaped using a computer software program called QuickTime and a built-in laptop camera, and this material was used for the stimulated recall session (see Section 3.4.2.), which immediately followed. This selective administration was to reduce the task load for the rest to encourage more people to participate in the study. Table 7 shows the distribution of the assigned task for the 20 participants. Participants assigned to the assessment tasks were relatively evenly distributed among the four recruiting groups.

Table 7. Overview of task assignment distribution for the stimulated recall subset (*n = 20*)

Task assignment	# of assigned participants				
	G1	G2	G3	G4	Total
ILST version A, biology	1	2	1	1	5
ILST version B, business	1	1	1	2	5
ILST total	2	3	2	3	10
MMST version A, biology	1	2	1	1	5
MMST version B, business	1	1	2	1	5
MMST total	2	3	3	2	10

Note. G1 = oral communication classes in an intensive English program (low proficiency); G2 = ESL listening courses (moderate proficiency); G3 = ESL speaking courses for international teaching assistants (moderate proficiency); G4 = a graduate program in applied linguistics (high proficiency)

3.4.2. Stimulated Recall

Immediately after completing the assigned assessment task, the subset of 20 participants watched a videotaped performance of themselves doing the task, played on a laptop computer. Following the stimulated recall protocol (see Section 3.3.3 and Appendix F), participants were asked to report what they had been thinking during task performance, During the session, either the participants or I could elect to pause the video for discussion of their recollected thinking during the assessment task. To further help participants recollect, task materials given to them in the previous step (see Appendices B and D) were also provided. The entire procedure with each participant took between 10 and 30 minutes depending on how much participants could remember, and was audiotaped for an analysis of strategic competence participants used to complete the assigned assessment task (described in Section 3.6.4).

3.4.3. Authenticity Questionnaires

After the stimulated recall interview (the subset of 20 participants) or after the assigned assessment task (the remaining 73), participants completed authenticity questionnaires: 1) the situational authenticity questionnaire that asked them to compare the similarity of the assigned assessment task to real-life ones (see Section 3.3.4 and Appendix G) and 2) the interactional authenticity questionnaire that asked them to evaluate the involvement of the construct-relevant language ability in accomplishing the assigned task (see Section 3.3.5 and Appendix H). Participants rated the questionnaire items on a six-point Likert scale (refer to Sections 3.3.4 and 3.3.5 for details). When participants, especially those who had low English proficiency, asked, I elaborated the meaning of questionnaire items to

assist them in understanding what the items intended to ask and providing ratings accordingly. Afterwards, I asked them to orally explain each of their questionnaire responses, and these were audiotaped. This oral explanation stage played a supplementary role of double-checking the reliability of participant ratings. In the rare cases when participant explanations did not seem to be relevant to the intended purpose of corresponding questionnaire items, I elaborated what the items meant and provided time for participants to consider whether their responses were reflective of the questionnaire items. Only when participants decided to change their rating(s), did I allow them to do so. The two activities, the questionnaire and the follow-up oral explanation, took less than 40 minutes.

The 20 participants recruited in Spring 2013 completed only the situational authenticity questionnaire. The preliminary construct analysis from the Spring 2013 data guided the selection of interactional authenticity questionnaire items, so the interactional authenticity questionnaire was developed after the Spring 2013 administration (as explained in Section 3.3.5). Figure 8 summarizes the entire research procedure and data collected in each step for the Spring 2013 participants. They provided spoken responses to their assigned task, strategic behavior reports during a simulated recall interview, and ratings and follow-up oral explanations on the situational questionnaire.

Figure 8. An overview of research procedure and collected data for the Spring 2013 participants

Figure 9 below is a summary of the research procedure and collected data for the 73 participants recruited in Fall 2014. The 73 participants answered the two authenticity questionnaires immediately after completing their assigned assessment task.

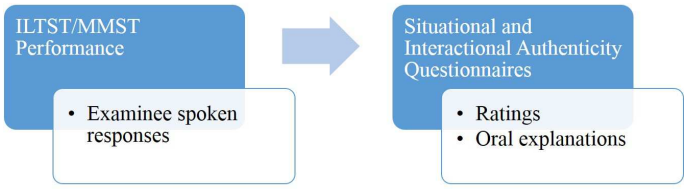


Figure 9. An overview of research procedure and collected data for the Fall 2014 participants

3.4.4. Scoring

Examinee spoken responses elicited from the ILST and the MMST were holistically rated for assigning a score of task performance. Three experienced raters, all of whom were certified raters of an oral proficiency test for international teaching assistants (ITAs)[6], scored the speech samples. The raters were first informed of task format, the content of the input lectures, and the rubrics, the *TOEFL® iBT Test Integrated Speaking Rubrics* (ETS, 2008). Next, each rater independently listened to the audio files of the 46 ILST and 47 MMST spoken responses, and assigned holistic scores from one (limited and unintelligible) to four (fulfilling and intelligible). The audio files contained no participant identifiers, and were presented to each rater in the same random order. The raters recorded all the scores on Excel spreadsheets, and then submitted those to me.

[6] Although the tasks of the ITA test are different from those of the current study, these raters can be considered having competence in rating speech samples in general.

3.5. Data

Four types of data were collected, each of which is described in detail in the following sub-sections: 1) transcribed spoken responses on the two assessment tasks, the ILST and the MMST, 2) assigned task scores, 3) transcribed strategic behavior reports from stimulated recall, and 4) authenticity questionnaire responses.

3.5.1. Transcribed Spoken Responses on the Two Assessment Tasks

Twenty-three spoken responses each on the ILST version A-biology and the ILST version B-business, and 23 spoken responses on the MMST version A-biology and 24 on the MMST version B-business were collected (refer to Table 6). I transcribed all of these 93 one-minute examinee spoken responses, and, in preparation for analysis, pruned the transcribed responses, excluding features of repair[7]. The goal here was to concentrate only on the primary meaning participants were making. The pruned transcriptions were saved in text files. These transcribed spoken responses are seen as evidence of participants' use of language knowledge and their language choices in response to the context created in the assigned assessment task.

3.5.2. Assigned Task Scores

Three raters independently assigned holistic scores, ranging from one to four, to the 46 examinee spoken responses on the ILST and the 47 responses on the MMST. A high degree of reliability was found among the ILST scores assigned by the three raters. The two-way mixed average measure intraclass correlation coefficient (ICC) was .852 with a 95% confidence interval from .746 to .916 ($F(45, 90)=7.701$, $p < .000$). A relatively high degree of

[7] Repair included repetition, rephrasing and false starts as defined in Freed (2000).

reliability was also found among the MMST scores assigned by the three raters. The two-way mixed average measure ICC was .751 with a 95% confidence interval from .578 to .856 ($F(46, 92)=4.618$, $p < .000$). Overall, there was relatively high inter-rater reliability.

In the circumstance where there was discrepancy, I chose the more frequent score on an examinee spoken response. For instance, when Raters 1 and 3 gave a score of three to a spoken response, but Rater 2 gave a score of one, a score of three was chosen as the assigned score for that response. In total, seven discrepancies out of 93 ratings by more than one point were found. All cases were due to lower ratings assigned by Rater 2, who was relatively stricter than the other two raters. When all three raters gave different ratings (e.g., two for Rater 1, one for Rater 2, and three for Rater 3), the median score was assigned (a score of two for the example above). In total, six complete disagreements among the three raters out of 93 ratings were found.

Table 8 below summarizes the descriptive statistics of the scores assigned to the ILST spoken responses.

Table 8. Descriptive statistics of the scores assigned to the ILST spoken responses (*n = 46*)

Assessment Task	N	Mean	Median	Mode	SD	Min	Max
ILST version A, biology	23	3.00	3	3	0.85	1	4
ILST version B, business	23	3.22	3	3	0.67	2	4
Total	46	3.11	3	3	0.77	1	4

A one-way between subjects ANOVA was conducted to evaluate the effect of lecture topics (biology vs. business) on the ILST scores, and there was no significant effect at the $p < .05$ level [$F(1, 44) = 0.923$, $p = 0.342$]. Different lecture topics did not affect the score of the ILST participants.

Table 9 below summarizes the descriptive statistics of the scores assigned to the MMST spoken responses.

Table 9. Descriptive statistics of the scores assigned to the MMST spoken responses (*n = 47*)

Assessment Task	N	Mean	Median	Mode	SD	Min	Max
MMST version A, biology	23	2.91	3	3	0.67	2	4
MMST version B, business	24	3.00	3	3	0.66	2	4
Total	47	2.96	3	3	0.66	2	4

A one-way between subjects ANOVA was conducted to evaluate the effect of lecture topics (biology vs. business) on the MMST scores, and there was no significant effect at the $p < .05$ level [$F(1, 45) = 0.202$, $p = 0.656$]. Different lecture topics did not also affect the score of the MMST participants.

3.5.3. Transcribed Strategic Behavior Reports from Stimulated Recall

Five strategic behavior reports each on the ILST version A-biology and the MMST version A-biology, and five strategic behavior reports each on the ILST version B-business and the MMST version B-business were collected (refer to Table 7). All 20 of the audiotaped stimulated recall data were fully transcribed. These transcribed strategic behavior reports were analyzed to find evidence of participants' use of strategic competence in response to the context created in the assigned assessment task.

3.5.4. Authenticity Questionnaire Responses

Two types of data were collected from all 93 participants for the situational authenticity questionnaire and from the 73 Fall 2014 participants for the interactional authenticity questionnaire (see Section 3.4.3). The first data type was a set of quantitative

ratings on a six-point Likert scale. Participants marked their ratings on a printed copy of the questionnaires, and I transferred them into Excel spreadsheets for analysis. The higher ratings indicate a higher degree of authenticity perceived by participants.

The second data type was a set of follow-up oral explanations for the choices of the assigned ratings to the questionnaire items. All of the 93 audio-recorded episodes of oral explanations for the situational authenticity responses and the 73 for the interactional authenticity responses were transcribed, and the transcriptions were saved in Word files. The transcribed oral explanations helped to explain the results of the questionnaire ratings.

3.6. Data Analysis

The data were analyzed to answer each of the research questions (see Section 2.8). In this section, I describe each of the data analyses conducted in the current study.

3.6.1. Grouping Based on the Task Scores

The third and fourth research questions focused on the participants' use of their language ability, consisting of language knowledge and strategic competence, during the performance of the two assessment tasks. Patterns of elicited language performance across varying degrees of proficiency level were also examined. This necessitated grouping participants based on their task scores. Based on the holistic scores assigned to examinee spoken responses (discussed in Section 3.5.2), participants were placed into three task score groups: low, moderate, or high. The ILST and the MMST are slightly different measures, as their channel and form of the input lecture is different, and possibly so is the relationship between the input and the response. Therefore, in a strict sense, the holistic scores from these

two measures cannot be treated as directly comparable. However, in the scoring rubrics used for assigning the ILST and the MMST scores, some evaluative features such as pronunciation, intonation, and grammatical accuracy are assumed to be independent from the format of the input lecture, so the holistic scores of the two tasks encapsulate some constant quality. Participants who received a score of one or two were placed in the low task score group, those who received a score of three were in the moderate group, and those with a score of four were placed into the high group. Table 10 presents the distribution of participants per task score in each of the assessment tasks.

Table 10. Distribution of participants across three task scores (*n = 93*)

Assessment Task	# of the assigned participants			
	Low	Moderate	High	Total
ILST version A, biology	6	10	7	23
ILST version B, business	3	12	8	23
ILST Total	9	22	15	46
MMST version A, biology	6	13	4	23
MMST version B, business	5	14	5	24
MMST Total	11	27	9	47

The distribution of participants across the three task scores is roughly similar between the tasks of different lecture topics within an assessment task type (e.g., the ILST version A, biology vs. the ILST version B, business) and also between the two assessment task types (i.e., the ILST vs. the MMST).

To check if a participant's initial recruiting group (from Group 1: the lowest to Group 4: the highest) corresponded to score-based grouping, I examined how participants from each of the four recruiting groups were placed into three task score groups per assessment task (see Table 11).

Table 11. Distribution of participants from the four recruiting groups across task scores (*n = 93*)

Assessment Task	# of the assigned participants											
	Low (*n = 20*)				Moderate (*n = 49*)				High (*n = 24*)			
	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4
ILST version A, biology	3	2	1	0	0	2	7	1	0	0	0	7
ILST version B, business	2	1	0	0	3	2	7	0	0	0	1	7
ILST Total	5	3	1	0	3	4	14	1	0	0	1	14
MMST version A, biology	3	1	2	0	1	4	6	2	0	0	0	4
MMST version B, business	3	0	2	0	2	4	7	1	0	0	0	5
MMST Total	6	1	4	0	3	8	13	3	0	0	0	9

Note. G1 = oral communication classes in an intensive English program (low proficiency); G2 = ESL listening courses (moderate proficiency); G3 = ESL speaking courses for international teaching assistants (moderate proficiency); G4 = a graduate program in applied linguistics (high proficiency)

Participants recruited from Group 1 (the lowest proficiency) attained mostly low scores on the tasks, whereas most of the participants recruited from Group 4 (the highest proficiency) attained high task scores. Those from Groups 2 and 3 attained mostly moderate scores. Therefore, it can be considered that the distribution of task scores was approximately alike in each assessment task.

3.6.2. Authenticity Questionnaire Analysis

To find out the perceived degree of 1) correspondence between the two assessment tasks and the target language use task and 2) language ability involved in accomplishing the two assessment tasks, I analyzed two types of data from the situational authenticity questionnaire (Appendix G) and the interactional authenticity questionnaire (Appendix H): 1) ratings on each questionnaire item and 2) oral explanations of the assigned ratings. First, I calculated the descriptive statistics (e.g., mean, standard deviation) of participants' ratings per task type (the ILST or the MMST) and treated the mean rating for each task type as the

total authenticity score. I ran a multivariate analysis of variance (MANOVA) to compare the degree of perceived situational authenticity between the ILST and the MMST, and a series of independent t-tests and MANOVA to compare the perceived interactional authenticity between the two tasks. I also used the transcripts of examinees' oral explanations to explain the quantitative findings of the questionnaire ratings.

3.6.3. Systemic Functional Linguistic Analysis

To obtain evidence about what language knowledge the ILST and the MMST elicited (Research Question 3), I analyzed the transcriptions of the pruned spoken responses on the two assessment tasks (refer to Section 3.5.1) regarding two SFL features. The analytical schemes were developed to investigate how participants expressed aspects of two metafunctions in SFL: *ideational* and *textual* (Halliday, 1978). An analysis of processes was conducted to analyze the language that participants used to express the *ideational* metafunctions, and a cohesion analysis to analyze the language representing the *textual* metafunctions. The following sub-sections describe each analysis in detail.

3.6.3.1. Analysis of processes

To analyze the language participants used to express the *ideational* metafunctions, or in other words, the content and activity of what occurred in the lecture, I analyzed *processes*, the activity or way of being in an event (e.g., *use*, *expect*) (Droga & Humphrey, 2002; Ravelli, 2000). *Processes* consist of seven types: 1) material, 2) mental, 3) verbal, 4) behavioral, 5) identifying relational, 6) attributive relational, and 7) existential. These elements offer a means for researchers to learn the *ideational* meaning of examinees' spoken

responses. The number of each *process* per task version (biology vs. business) and type (the ILST vs. the MMST) were calculated for transforming qualitative data into quantitative. The counts of *processes* served as the data for comparison between the two assessment task types (Research Question 3a) and versions (Research Question 3b). I ran a chi-squared test to compare the patterns of the *process* use between the two assessment task versions and types. Using the score-based grouping, explained in Section 3.6.1, differences across assigned task scores were also examined, using a three-way loglinear analysis.

3.6.3.2. Cohesion analysis

The analysis of the *textual* metafunctions (how the message is organized) was performed to answer the research question, 3c, regarding differences in the *cohesion* (Collerson, 1994) of spoken responses between the ILST and the MMST. I paid specific attention to pronouns used as an exophoric reference (something outside of the text itself), due to the construct-relevance of this linguistic device. A *cohesion* analysis demonstrates connections between different parts of the message. In the current study, the pronoun analysis served as a means to discover participants' organizational choices made in their spoken responses. The number of each type of pronoun (e.g., I, you, we, your, our, us) used as an exophoric reference per task (the ILST vs. the MMST) was calculated for transforming the qualitative data into the quantitative. The total counts of pronouns as an exophoric reference served as the data for comparison between the two assessment tasks. I ran a chi-squared test to compare the use of exophoric pronouns between the ILST and the MMST. Using the score-based grouping, explained in Section 3.6.1, differences across the assigned task scores were also examined, using a three-way loglinear analysis.

The following table outlines all the SFL aspects used in the data analysis.

Table 12. SFL analytical features

Metafunction	Features		Example
Ideational	Transitivity System	Process	*use*, *choose*
Textual	Pronouns	Exophoric reference	*I*, you, *we*, *your*, *our, us*

3.6.4. Strategic Competence Analysis

To investigate how strategic competence was involved in completing the two assessment tasks (Research Question 4), I analyzed 20 participants' strategic behavior reports collected from simulated recall sessions. I coded the transcripts of the strategic behavior reports according to a modified version of Swain et al.'s (2009) theoretically grounded coding scheme (see Appendix I). Modifications include strategies relevant to the context of this study only. For instance, the fact that the ILST and the MMST do not involve reading resulted in removing the strategy of "recalling the text" from the coding scheme. In addition, participants from my preliminary analysis of Spring 2013 data did not report some of the communication strategies such as "simplifying the message" during the stimulated recall, and those unreported strategies were eliminated from the coding scheme.

The modified coding scheme consists of the following five categories of strategies: 1) approach, 2) communication, 3) cognitive, 4) metacognitive, and 5) affective. The approach strategies refer to what examinees do to orient themselves to the task such as thinking about the meaning of the question given in the assessment task. The communication strategies involve conscious plans for solving a linguistic problem in order to reach a communicative goal such as reviewing notes to remember/formulate what to say. The cognitive strategies entail the manipulation of the target language to understand and produce language such as

anticipating the content of the input lecture in the assessment task. The metacognitive strategies include seeking a goal for completing the assigned task, planning the parts, sequence, or main ideas to be expressed in spoken responses, and evaluating language production while speaking. The affective strategies concern self-talk or mental control over affect such as examines justifying their performance.

The analysis identified reported behavior indicative of patterns of strategic competence used to complete the ILST and the MMST, quantitatively transformed to raw counts of different types of strategies per task type. The descriptive statistics served as the data for comparison between the two assessment tasks. Since the number of participants in each assessment task group was small, only descriptive statistics were used for the analysis. Stimulated recall quotes corresponding to each strategy type provided supplementary information to interpret the quantitative findings. Interesting findings from the comparative analysis between the two assessment task groups were further examined across the assigned task scores described in Section 3.6.1.

3.7. Chapter Summary

The perceived authenticity analysis of the current study employed the embedded data model (Creswell & Plano Clark, 2007), and the analysis of language ability elicited during task performance used a qualitative research design with data transformation. This two-part research design was planned to examine the authenticity of the two assessment tasks, the ILST and the MMST from students' perception and evidence of their language ability elicited by the tasks.

Ninety-three international students with varying degrees of English proficiency participated in the study. They were randomly assigned to doing either the ILST or the MMST. After the assigned task, participants responded to authenticity questionnaires, and a subset of 20 did stimulated recall interviews. Ratings on the questionnaires and analysis of oral explanations on the questionnaires were used to provide evidence of the participants' perception of the authenticity of the two assessment tasks. Spoken responses on the tasks and stimulated recall reports were used to analyze how participants' language knowledge and strategic competence were actually involved in completing the assessment tasks. Differences between the two assessment tasks and among the assigned task scores were also examined. Table 13 below summarizes types of data and analytic methods used for answering each research question.

Table 13. Summary of data and analyses used for answering each research question

Research Question	Data	N	Analysis
1. To what degree do examinees, who are students in the target language use (TLU) context, perceive each of the two assessment tasks—the ILST and the MMST—as corresponding the TLU task, explaining/informing tasks in the academic domain? How different is this perceived situational authenticity between the two assessment tasks?	Situational authenticity questionnaire ratings and transcribed oral explanations	ILST: 46 MMST: 47	Descriptive statistics analysis Multivariate analysis of variance (MANOVA)
2. To what degree do examinees consider the involvement of language ability, defined in the construct, to be high in accomplishing each of the two assessment tasks? How different	Interactional authenticity questionnaire ratings and transcribed oral explanations	ILST: 36 MMST: 37	Descriptive statistics analysis Independent-samples t-test Multivariate analysis of variance

Table 13. (continued)

is this perceived interactional authenticity between the two assessment tasks?			(MANOVA)
3. What language knowledge do the two assessment tasks—the ILST and the MMST—elicit in the examinees' spoken response? How different is this elicited language knowledge between the tasks and across assigned task scores?			
3a. How different is the use of processes between the two assessment tasks? Are there differences affected by task scores?	Transcribed examinee spoken responses Task scores	Biology: 46 Business: 47	Analysis of processes Descriptive statistics Chi-squared test Three-way loglinear analysis
3b. How different is the use of processes between the task versions, biology vs. business?	Transcribed examinee spoken responses	ILST: 46 MMST: 47	Analysis of processes Descriptive statistics
3c. How different is the use of exophoric reference between the two assessment tasks? Are there any notable differences across assigned task scores?	Transcribed examinee spoken responses Task scores	ILST: 46 MMST: 47	Cohesion analysis Descriptive statistics Chi-squared test Three-way loglinear analysis
4. How is strategic competence involved in completing the two types of assessment tasks? Are there any notable differences between the two task types? Do these differences vary across assigned task scores?	Transcribed strategic behavior reports Task scores	ILST: 10 MMST: 10	Stimulated recall analysis Descriptive statistics

Findings from analyses using these data and analytic approaches will be presented in the following chapter.

CHAPTER 4. RESULTS

This chapter provides answers to the research questions using the findings of multiple analyses outlined in the previous chapter. Specifically, the discussion in this chapter covers 1) perceived *situational authenticity* of the two assessment tasks, the Integrated Listening-Speaking Task (ILST) and the Multimedia-Mediated Speaking Task (MMST), 2) perceived *interactional authenticity* of the assessment tasks, 3) elicited language knowledge manifested in student task performance, and 4) strategic competence involved during student task performance. The data analyzed to answer the four research questions consisted of ratings on the situational and interactional authenticity questionnaire, and transcripts of 1) the follow-up interviews regarding student perception of situational and interactional authenticity, 2) student spoken responses to the assigned assessment task, and 3) stimulated recall interviews. The presentation and discussion of the results in this chapter are structured around each individual research question.

4.1. Perceived Situational Authenticity of the Assessment Tasks

This section reports on the results of the first research question: To what degree do the examinees, who are students in the target language use (TLU) context, perceive the two assessment tasks as corresponding the TLU task, in other words, the perceived *situational authenticity* (Bachman, 1991)? The data consist of student ratings on the situational authenticity questionnaire consisting of ten items (Appendix G), completed by students who performed either the ILST or the MMST. Transcripts of the follow-up interviews about the

reasons for assigning certain ratings provide supplementary information for understanding the students' perceptions of the situational authenticity of each assessment task.

In the first two sub-sections, I will present the descriptive statistics showing how students perceived the characteristics of two assessment tasks relative to those of the TLU task, an explaining/informing task in the academic domain: first, the integrated listening-speaking task (ILST) and second, the multimedia-mediated speaking task (MMST). To assist the interpretations, quotes from the interview transcripts that were particularly revealing of student opinions are also provided. In the third sub-section, I will discuss how the ratings were statistically compared between the two groups, the ILST and the MMST examinees, using multivariate analysis of variance (MANOVA). The section will end with a summary.

4.1.1. Integrated Listening-Speaking Task (ILST)

Forty-six students performed the integrated listening-speaking task (ILST), and responded to the ten items of the situational authenticity questionnaire on a six-point Likert scale. Each rating indicates the degree of similarity between the ILST and the TLU task on this scale (1 = very different; 6 = very close). Table 14 below presents the item-level descriptive statistics of the ratings from the group of the students who were assigned to the ILST. The mean of the item statistics denotes the average rating across all ten questionnaire items, regarded as the overall degree of perceived situational authenticity of the ILST, which can be interpreted on the same scale as the item statistics.

Table 14. Descriptive statistics for the ILST item-level ratings of situational authenticity ($N = 46$)

	Situational Authenticity Questionnaire Items	Mean	Median	Mode	SD	Min	Max
1	Task rubrics	4.80	5.00	6	1.20	1	6
2	Setting	4.52	5.00	5	1.33	2	6
3	Participant	4.96	5.00	6	1.23	2	6
4	Functional characteristics	4.93	5.00	6	1.06	2	6
5	Format	4.41	5.00	5	1.51	1	6
6	Organizational characteristics	4.80	5.00	5	1.22	1	6
7	Sociolinguistic characteristics	4.96	5.00	6	1.15	2	6
8	Problem identification	4.39	5.00	6	1.53	1	6
9	Relationship between input and response*	4.42	5.00	5	1.38	1	6
10	Genre	4.48	5.00	5	1.30	1	6
	Overall*	4.74	4.70	4.70	0.67	3.30	6.00

* Item 9 (relationship between input and response) was added at the time of the Fall 2014 administration; therefore, the statistics are based on the ratings from 36 students only.

The mean of the item-level ratings was treated as indicating the students' overall perception of situational authenticity. The scale of situational authenticity of the ILST had relatively high reliability, Cronbach's $\alpha = .72$. The overall perception was relatively high ($M = 4.74$, $SD = 0.67$), and empirically supports the popular belief in the authentic nature of integrated assessment tasks (Weir, 1990; Plakans, 2013).

The item-level mean ratings range from moderately close to the target language use (TLU) task (e.g., $M = 4.39$ for item 8) to closer to the TLU task (e.g., $M = 4.96$ for items 3 and 7). Task characteristics that students perceived highly close to the TLU task are on participant (item 3) and sociolinguistic characteristics (item 7) ($M = 4.96$ for both). Students who performed the ILST viewed the professor in the input lecture as highly similar to one they might encounter in their real-life academic context. It seems that a still picture of the professor with aural information was sufficient for students to regard him or her as similar to a professor in real life. For instance, Student 1.6, one of the 19 students who gave item 3 a

rating of six, reported, "[s]ometimes the professor don't[8] use blackboard or something, just speak using words, you need to take notes and figure out." He pointed out a frequent occurrence of the aural-only medium in academic lectures. Students in the ILST group also indicated that the task prompted them to speak similarly to the way they would in an academic situation (item 7, sociolinguistic characteristics).

In addition, they considered the goal targeted by the ILST (i.e., explaining/informing about an academic concept) highly close to that of the TLU task ($M = 4.93$ for item 4, functional characteristics). For example, Student 2.8, one of the 18 students who assigned item 4 a rating of six, reported:

> … when the lecturer wanted to make sure if the students really understand what he explained, and he may ask the students such questions… The professor from our senior class, sometimes ...yeah, especially when the professors want to interact with students, they may ask such kind of questions.

This quote suggests that the targeted goal of the task, explaining/informing, is often asked for students to accomplish in academic classes.

Students in the ILST group also regarded the information given in the ILST highly similar to what they would receive in the TLU task in order to accomplish the assigned task (item 1, task rubrics), and the language used in the input lecture and the corresponding real-life one highly alike (item 6, organizational characteristics) ($M = 4.80$ for both items). For instance, Student 1.6, one of the 14 students who item 6 gave a rating of six, provided a supporting remark that "the guy [the professor in the input lecture] has a good pronunciation, vocabulary, grammar, very clear, we can understand. He's like a professor, yeah."

[8] Any grammatical errors found in student quotes here and afterwards are verbatim transcripts, not misprints.

4.1.2. Multimedia-Mediated Speaking Task (MMST)

Forty-seven students performed the multimedia-mediated speaking task (MMST), and responded to the ten items of the situational authenticity questionnaire on a six-point Likert scale. Each rating indicates the degree of similarity between the MMST and the TLU task on this scale (1 = very different; 6 = very close). Table 15 below presents the item-level descriptive statistics of the ratings from the group of the students who were assigned to the MMST. The mean of the item statistics denotes the average rating across all ten questionnaire items, regarded as the overall degree of perceived situational authenticity of the MMST, which can be interpreted on the same scale as the item statistics.

Table 15. Descriptive statistics for the MMST item-level ratings of situational authenticity ($N = 47$)

	Situational Authenticity Questionnaire Items	Mean	Median	Mode	SD	Min	Max
1	Task rubrics	4.91	5.00	6	1.21	1	6
2	Setting	4.70	5.00	6	1.23	1	6
3	Participant	4.87	5.00	6	1.26	1	6
4	Functional characteristics	4.96	5.00	6	1.14	2	6
5	Format	4.30	5.00	6	1.64	1	6
6	Organizational characteristics	5.02	5.00	6	1.05	2	6
7	Sociolinguistic characteristics	4.62	5.00	6	1.26	2	6
8	Problem identification	4.77	5.00	5	1.03	2	6
9	Relationship between input and response*	4.05	4.00	4	1.22	1	6
10	Genre	4.64	5.00	4	1.21	1	6
	Overall*	4.58	4.70	4.10	0.76	2.90	5.80

* Item 9 (relationship between input and response) was added at the time of the Fall 2014 administration; therefore, the statistics are based on the ratings from 36 students only.

As in the analysis of the ILST data, the mean of the item-level ratings was treated as indicating the students' overall perception of situational authenticity. The scale of situational authenticity of the MMST had high reliability, Cronbach's $\alpha = .83$. The overall perception of situational authenticity is moderate ($M = 4.58$, $SD = 0.76$). Compared to the students assigned to performing the ILST, students in both groups had similar perceptions ($M = 4.74$, $SD = 0.67$

for the ILST; $M = 4.58$, $SD = 0.76$ for the MMST). Based on the perception by the students of the current study, it appears that using multimodal input did not increase the authenticity of the assessment tasks, contrary to what some language testing researchers have proposed (e.g., Chapelle and Douglas, 2006; Douglas and Hegelheimer, 2007).

The MMST item-level mean ratings range from moderately close to the target language use (TLU) task (e.g., $M = 4.05$ for item 9) to highly close (e.g., $M = 5.02$ for item 6), and overall, are roughly the same as or slightly higher than those from the ILST group, except those on item 7, sociolinguistic characteristics ($M = 4.96$ for the ILST vs. $M = 4.62$ for the MMST) and item 9, relationship between input and response ($M = 4.42$ for the ILST vs. $M = 4.05$ for the MMST). As in the ILST group, students in the MMST group also perceived a relatively high degree of situational authenticity on the following aspects: item 1, task rubrics ($M = 4.91$); item 3, participant ($M = 4.87$); item 4, functional characteristics ($M = 4.96$); and item 6, organizational characteristics ($M = 5.02$). In addition, they viewed the characteristic of requiring a problem-solving task (i.e., explaining aspects of an academic topic learned from a lecture) as highly similar to that of what they would perform in real life ($M = 4.77$ for item 8, problem identification). For example, Student 1.11, one of the 13 students who gave Item 8 a rating of six, reported in the follow-up interview:

> I think the task to answer what I learned in the lecture is so similar to the real life because … when the teachers are giving a lecture in the real life, during the lectures, … the teachers ask for the students if it does make sense. So probably the students should demonstrate if they are understanding the content and sometimes to ask some questions and sometimes to comment. So it is so close.

In this quote, the student reiterated a close similarity between the MMST and the classroom task. In addition, Student 3.10, who also assigned a rating of six, pointed out what the MMST required is similar to outside-classroom tasks with peers, saying "because there might some student, my friend may ask me some questions because they didn't really understand that topic. I have to talk about that topic again to teach them like that. I have to use this kind of topic, task."

Additionally, students in the MMST group considered the context, simulated by the MMST, highly similar to that of their academic context (M = 4.70 for item 2, setting), as suggested by the following quote from Student 3.10, one of the 15 students who gave item 2 a rating of six, "[t]his PPT [PowerPoint slides used in the input lecture] helps me to feel like I am in classroom." It appears that content visuals included in the video lecture enhanced the authenticity of the task setting.

Interestingly, the MMST student ratings on types of language elicited by the assigned task (M = 4.62 for item 7, sociolinguistic characteristics) were lower than those from the ILST group (M = 4.96). Students from the MMST group also considered the relationship between the input lecture and their spoken response as only moderately authentic (M = 4.05 for item 9), compared to the higher mean rating of 4.42 by the ILST students.

4.1.3. Statistical Comparison Between the ILST and the MMST

Sub-sections 4.1.1 and 4.1.2 discussed the perceived situational authenticity of the integrated listening-speaking task (ILST) and the multimedia-mediated speaking task (MMST), respectively, and particularly in Sub-section 4.1.2, a comparison between the two assessment tasks was also made using item-level descriptive statistics. In this sub-section, a

statistical comparison between the two assessment tasks on the perceived situational authenticity is made, based on the results of one-way multivariate analysis of variance (MANOVA).

The ratings on the ten situational questionnaire items from 36 students in the ILST group and 37 students in the MMST group were statistically compared (73 students in total). As item 9 (relationship between input and response) was added at the time of the Fall 2014 administration, ten students in the ILST and the MMST group, respectively, from the Spring 2013 administration were excluded from the analysis due to missing data. Using Wilk's lambda, there was not a significant difference between the ILST and the MMST in the ratings on the ten situational questionnaire items collectively, $\Lambda = 0.88$, $F(10, 62) = 0.82$, $p = .614$.

Separate univariate ANOVAs on the questionnaire ratings also revealed non-significant group differences on each of the ten aspects of the perceived situational authenticity: 1) task rubrics, $F(1, 71) = 0.32$, $p = .574$, 2) setting, $F(1, 71) = 0.63$, $p = .429$, 3) participant, $F(1, 71) = 0.59$, $p = .443$, 4) functional characteristics, $F(1, 71) = 0.10$, $p = .757$, 5) format, $F(1, 71) = 1.07$, $p = .305$, 6) organizational characteristics, $F(1, 71) = 0.11$, $p = .739$, 7) sociolinguistic characteristics, $F(1, 71) = 2.47$, $p = .121$, 8) problem identification, $F(1, 71) = 0.52$, $p = .471$, 9) relationship between input and response, $F(1, 71) = 1.41$, $p = .239$, and 10) genre, $F(1, 71) = 0.25$, $p = .616$. Although students in the two groups, the ILST and the MMST, showed a seemingly different degree of situational authenticity on a few aspects based on the item-level comparisons of their questionnaire ratings, it appears that they, in fact, considered the degree of the situational authenticity of both assessment tasks similar, both overall and for each authenticity aspect. Alternatively, statistical non-significance between the groups could have been due to a ceiling effect in participant ratings.

The distributions of ratings on most of the questionnaire items were left-skewed with the medians of five and the modes of five or six on a six-point Likert scale, and this negative skewness occurred because the questionnaire instrument may have not properly measured variance in participant perception of situational authenticity at the upper end of the scale. More sensitive questionnaire would have captured possibly different degrees of perception between the two groups of participants.

To include more cases of complete data, other analyses excluding the ratings on item 9 (relationship between input and response) were conducted, so the ratings on the nine situational questionnaire items from 46 students in the ILST group and 47 students in the MMST group were statistically compared (93 students in total). Using Wilk's lambda, there was not a significant difference between the ILST and the MMST in the ratings on the nine situational questionnaire items collectively, $\Lambda = 0.93$, $F(9, 83) = 0.74$, $p = .67$. Separate univariate ANOVAs on the questionnaire ratings also revealed non-significant group differences on each of the nine aspects of the perceived situational authenticity: 1) task rubrics, $F(1, 91) = 0.20$, $p = .660$, 2) setting, $F(1, 91) = 0.46$, $p = .499$, 3) participant, $F(1, 91) = 0.11$, $p = .745$, 4) functional characteristics, $F(1, 91) = 0.01$, $p = .921$, 5) format, $F(1, 91) = 0.12$, $p = .726$, 6) organizational characteristics, $F(1, 91) = 0.84$, $p = .361$, 7) sociolinguistic characteristics, $F(1, 91) = 1.83$, $p = .179$, 8) problem identification, $F(1, 91) = 1.94$, $p = .167$, and 10) genre, $F(1, 91) = 0.38$, $p = .539$. These findings conform to those based on the ratings on the ten questionnaire items.

4.1.4. Section Summary

To address the first research question, the ratings of *situational authenticity* assigned by students who took either of the two assessment tasks, the ILST and the MMST, were analyzed. Students in the ILST group perceived a relatively high degree of situational authenticity overall ($M = 4.74$, $SD = 0.67$) on a six-point Likert scale. The item-level mean ratings ranged from moderately close to the target language use (TLU) task, (e.g., 4.39 for item 8, problem identification) to closer to the TLU task (e.g., 4.96 for item 7, sociolinguistic characteristics). These student perceptions empirically support the widely accepted notion of authenticity in integrated assessment tasks (Weir, 1990; Plakans, 2013).

In addition, students perceived the overall degree of situational authenticity of the MMST ($M = 4.58$, $SD = 0.76$) as similar to that of the ILST. The one-way multivariate analysis of variance (MANOVA) statistically confirmed the non-significant difference of perceived situational authenticity between the ILST and the MMST, $\Lambda = 0.88$, $F(10, 62) = 0.82$, $p = .614$. The item-level mean ratings assigned by students in the MMST group ranged from moderately close to the target language use (TLU) task (4.05 for item 9, relationship between input and response) to highly close (5.02 for item 6, organizational characteristics). Although students in the two groups showed a seemingly different degree of situational authenticity on a few aspects based on the item-level statistics (e.g., $M = 4.39$ for the ILST vs. $M = 4.77$ for the MMST on item 8, problem identification; $M = 4.52$ for the ILST and $M = 4.70$ for the MMST on item 2, setting), separate univariate ANOVAs on the questionnaire ratings revealed non-significant group differences on each of the ten aspects of the perceived situational authenticity. Adding multiple modes to the input lectures in the form of video did not enhance the situational authenticity of the integrated tasks from the students' perception.

This finding differs from what some language testing researchers have proposed (e.g., Chapelle and Douglas, 2006; Douglas and Hegelheimer, 2007) or resulted from a possibly less sensitive scale on the situational authenticity questionnaire that may have not measured subtle differences in participant perception.

4.2. Perceived Interactional Authenticity of the Assessment Tasks

This section reports on the results of the second research question: To what degree do the examinees, who are students in the target language use (TLU) context, consider the involvement of language ability, as defined in the construct, to be high in accomplishing each of the two assessment tasks? In other words, a degree of the perceived *interactional authenticity* of the two tasks was investigated. The data consist of student ratings on the interactional authenticity questionnaire consisting of two parts, language knowledge (items 1 – 6) and strategic competence (items 7 – 14) (Appendix H). These items were completed by students who took either the ILST or the MMST. Transcripts of the follow-up interviews about the reasons for assigning certain ratings provide supplementary information for understanding the students' perceptions of interactional authenticity of each assessment task.

In the first two sub-sections, I will present the descriptive statistics showing how students perceived to what extent their language ability was involved in accomplishing the assigned task, first, the integrated listening-speaking task (ILST) and second, the multimedia-mediated speaking task (MMST). To assist the interpretations, quotes from the interview transcripts that were particularly revealing of student opinions are also provided. In the third sub-section, I will discuss how the ratings were compared between the two groups, the ILST and the MMST examinees, using independent *t*-tests and multivariate analysis of variance

(MANOVA), and across the three assigned scores (low, moderate, and high), using visual presentations of line graphs. The section will end with a summary.

4.2.1. Integrated Listening-Speaking Task (ILST)

Thirty-six students[9] among the 46 who performed the integrated listening-speaking task (ILST) responded to the fourteen items of the interactional authenticity questionnaire on a six-point Likert scale. Each rating indicates the extent of involvement of students' language ability, consisting of language knowledge and strategic competence, in accomplishing the ILST (1 = not at all; 6 = a lot). The following two parts present the students' perceptions of first, their degree of language knowledge and second, their degree of strategic competence used during the ILST performance.

4.2.1.1. Language knowledge

Items 1 to 6 in the interactional authenticity questionnaire (Appendix H) concern the types of language knowledge defined in the task construct. Among the six, items 2, 3, and 4 regard the linguistic aspects related to topical differences in the input lectures (i.e., biology vs. business lecture), and accordingly, are excluded from the current analysis[10]. Items 1, 5, and 6 represent the core linguistic features, as defined in the construct, regardless of topical differences between the input lectures. Table 16 below presents the item-level descriptive statistics of the three ratings from the group of students who were assigned to the ILST.

[9] The interactional authenticity questionnaire was developed based on the preliminary findings from the Spring 2013 administration, so only the students of the Fall 2014 administration completed this questionnaire.
[10] The analysis of the ratings on these items will be reported in Section 4.3.2.

Table 16. Descriptive statistics for the ILST item-level ratings of interactional authenticity: Language knowledge ($N = 36$)

	Interactional Authenticity Questionnaire Items: Language Knowledge	Mean	Median	Mode	SD	Min	Max
1	Relational processes: Identifying	4.78	5.00	6	1.31	1	6
5	Existential processes	3.83	4.00	6	1.75	1	6
6	Pronouns: Exophoric reference	2.58	2.00	1	1.81	1	6

Note. Identifying relational processes were presented in the questionnaire as "language to give the entity a definite identity (e.g., (*the entity*) is (*a definite identity*); (*the entity*) means (*a definite identity*)"; existential processes as "language to state that something exists (e.g., there are…)"; exophoric pronouns as "language to refer to someone outside the context of your spoken response (e.g., we, you, your…)."

The item-level mean ratings range from a low (e.g., $M = 2.58$ for item 6) to high degree of involvement (e.g., $M = 4.78$ for item 1). According to student perceptions, knowledge of identifying relational processes (item 1), language for giving the entity in question a definite identity such as "X is Y," was often used during the ILST performance ($M = 4.78$). It seems that the task goal that requires students to define two aspects of an academic concept elicited the frequent use of students' knowledge of identifying relational processes. For instance, Student 3.11, one of the 19 students who assigned item 1 a rating of six, reported, "[b]ecause I need to use the definition of, for example, two property that people care … I need to use the definition of kind of this, so, six."

On the other hand, students reported that they moderately used their knowledge of existential processes (item 5), language for stating that something exists such as "there are X and Y," during their task performance ($M = 3.83$). This finding meets my expectation, as the ILST requires only one classification of either definitions of a tool (biology) or factors of product quality (business), which is usually manifested by the use of existential processes.

However, students perceived a less frequent use of their knowledge of pronouns as exophoric reference (e.g., "we/our/us" referring to the speaker and the imagined audience,

"you/your/you" referring to the imagined audience, and "I" referring to the speaker) (M = 2.58 for item 6). It appears that a still picture of a professor in the classroom, included in the ILST, did not provide sufficient information to create a sense of audience.

To examine the degree of interactional authenticity regarding language knowledge across three sub-groups of students according to their assigned task scores (low, moderate or high), the item-level mean ratings and standard deviations per score group were calculated (refer to Table 17).

Table 17. Descriptive statistics for the ILST item-level ratings of interactional authenticity per assigned task score: Language knowledge (N = 36)

	Interactional Authenticity Questionnaire Items: Language Knowledge	Mean (Standard Deviation)		
		Low (n = 6)	Moderate (n = 18)	High (n = 12)
1	Relational processes: Identifying	3.33 (1.51)	5.06 (0.80)	5.08 (1.44)
5	Existential processes	2.33 (1.03)	3.67 (1.75)	4.83 (1.47)
6	Pronouns: Exophoric reference	1.83 (1.60)	2.39 (1.46)	3.25 (2.26)

Students in these three sub-groups perceived the extent of interactional authenticity on the first two items as different. First, students perceived different degrees regarding how much knowledge of identifying relational processes (item 1) was used during their ILST performance. Specifically, six students in the low score group showed a lower degree of involved knowledge of this kind (M = 3.33) than those in the moderate (M = 5.06) and high groups (M = 5.08), depicted as the blue line in Figure 10.

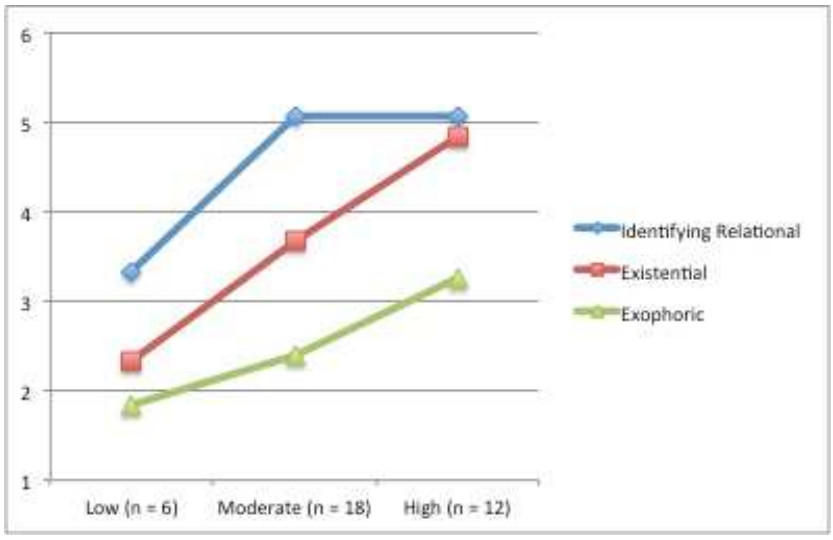


Figure 10. The ILST mean ratings of interactional authenticity on the language knowledge items among the three score groups ($N = 36$)

Six students in the low score group showed again a lower degree of involved knowledge of existential processes ($M = 2.33$); on the other hand, 12 students in the high score group showed a higher degree ($M = 4.83$) (refer to the red line in Figure 10). However, student perceptions did not differ greatly in the involved knowledge of pronouns as exophoric reference across the three score groups (see the green line in Figure 10).

4.2.1.2. Strategic competence

Items 7 to 14 in the interactional authenticity questionnaire (Appendix H) concern the types of strategic competence defined in the task construct. Table 18 below presents the item-level descriptive statistics of the ratings from the group of the students who were assigned to the ILST. The mean of the item statistics denotes the average rating across all eight questionnaire items, regarded as the overall degree of perceived interactional authenticity of the ILST regarding strategic competence. This figure allows an interpretation on the same scale as the item statistics.

Table 18. Descriptive statistics for the ILST item-level ratings of interactional authenticity: Strategic competence ($N = 36$)

	Interactional Authenticity Questionnaire Items: Strategic competence	Mean	Median	Mode	SD	Min	Max
7	Setting goals	4.78	5.00	6	1.29	2	6
8	Anticipating the content	4.11	4.00	4	1.55	1	6
9	Using imagery	3.33	3.00	1	1.90	1	6
10	Using notes	5.14	6.00	6	1.29	1	6
11	Reviewing notes	4.94	5.00	6	1.37	1	6
12	Planning	4.64	5.00	6	1.42	1	6
13	Referring to notes	5.00	5.00	5	1.04	2	6
14	Evaluating language production	3.56	4.00	4	1.28	1	6
	Overall	4.44	4.38	4.38	0.90	1.63	5.88

The mean of the item-level mean ratings was treated as indicating the students' overall perception of interactional authenticity regarding strategic competence. The scale of interactional authenticity (strategic competence) of the ILST had high reliability, Cronbach's $\alpha = .79$. The overall perception is moderate ($M = 4.44$, $SD = 0.90$).

The item-level mean ratings range from a moderate (e.g., $M = 3.33$ for item 9) to high degree of involvement (e.g., $M = 5.14$ for item 10). Overall, students in the ILST group reported a high degree of involvement regarding the following three aspects of strategic competence: 1) using notes to organize or remember information while listening to the input lecture ($M = 5.14$, item 10), 2) reviewing notes to remember or formulate what to say during response preparation ($M = 4.94$, item 11), and 3) referring to notes to remember or formulate what to say while speaking ($M = 5.00$, item 13). Students perceived a high degree of using their notes throughout all phases of the assigned task, from listening to the input lecture to preparing and providing a one-minute spoken response. For instance, Student 1.6, one of the 11 students who assigned items 10, 11, and 13 all a rating of six, reported a frequent use of strategies related to note taking and its usefulness in the following quote:

> Of course, if I take notes, I look for my notes to remember about the lecture, and organize my ideas to answer the question. Also this helped me a lot … I didn't write nothing, but my notes is very clear, with a good sequencing. So I just review my notes and prepare my answer. So I use this a lot ... Also I use because the lecture is not so big lecture, so I take notes and look again my notes. Check, organize my ideas in my mind.

Moreover, students in the ILST group perceived a relatively high level of using the strategy of setting goals for completing the assigned assessment task (item 7) ($M = 4.78$). For example, Student 3.19, one of the 15 students who gave item 7 a rating of six, said, "I think I have set the goal to explain that I have, I want to explain the concept and the two concepts and the compilation and their importance. I wanted to do that, so I chose six." He reported that he set a goal of explaining a concept, its two factors, and their relationship to accomplish the given task.

To examine the degree of interactional authenticity regarding strategic competence across three sub-groups according to the assigned task scores (low, moderate or high), the item-level mean ratings and standard deviations per score group were calculated (refer to Table 19). The eight aspects of strategic competence in the interactional authenticity questionnaire can be categorized into three types of strategies, derived from Swain et al. (2009), 1) communication strategies (items 11 and 13), 2) cognitive strategies (items 8, 9, and 10), and 3) metacognitive strategies (items 7, 12, and 14), and are labeled as such in the table.

Table 19. Descriptive statistics for the ILST item-level ratings of interactional authenticity per assigned task score: Strategic competence ($N = 36$)

	Interactional Authenticity Questionnaire Items: Strategic Competence	Mean (Standard Deviation)		
		Low ($n = 6$)	Moderate ($n = 18$)	High ($n = 12$)
	Communication Strategies			
11	Reviewing notes	3.83 (1.84)	5.06 (1.35)	5.33 (0.89)
13	Referring to notes	4.33 (1.21)	4.83 (1.04)	5.58 (0.67)
	Cognitive Strategies			
8	Anticipating the content	3.50 (1.64)	4.06 (1.43)	4.50 (1.68)
9	Using imagery	3.17 (2.14)	3.39 (1.72)	3.33 (2.19)
10	Using notes	4.33 (1.86)	5.06 (1.35)	5.67 (0.49)
	Metacognitive Strategies			
7	Setting goals	3.50 (1.34)	4.44 (1.10)	5.92 (0.29)
12	Planning	3.33 (1.75)	4.56 (1.34)	5.42 (0.79)
14	Evaluating language production	3.00 (0.89)	3.39 (1.20)	4.08 (1.44)
	Overall	3.63 (1.10)	4.35 (0.77)	5.00 (0.66)

The overall degree of perceived interactional authenticity regarding strategic competence showed an increasing trend across the three assigned task scores ($M = 3.63$ for low; $M = 4.35$ for moderate; $M = 5.00$ for high), depicted as the blue line in the next three figures (Figures 11, 12 and 13). This increasing tendency across the three scores slightly varies per sub-strategy.

Communication strategies, the first category of strategic competence, are defined as conscious plans for solving a linguistic problem in order to reach a communicative goal. The mean ratings of two strategy types under this category were higher than those of the overall (the blue line in Figure 11) across all three assigned task scores. Students across the assigned task scores perceived a relatively high level of using the strategies of reviewing and referring to notes to remember/formulate what to say. Noticeably, the mean rating of the reviewing notes strategy showed a sharp increase from the low ($M = 3.83$) to the moderate score group

(M = 5.06). Students in the ILST group perceived generally a frequent use of the communication strategies during their task performance.

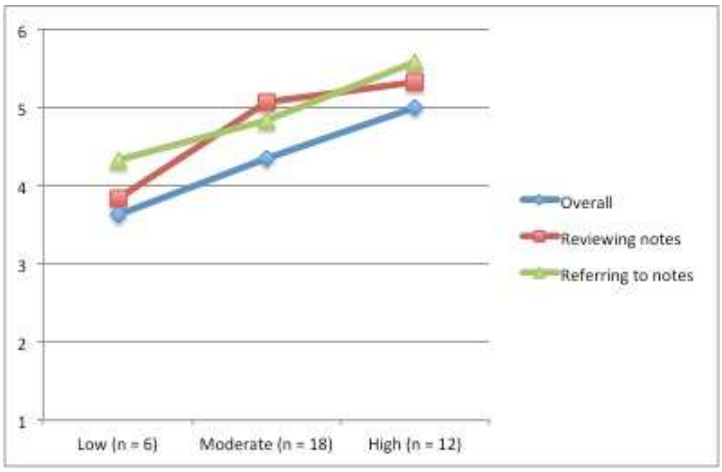


Figure 11. The ILST mean ratings of interactional authenticity on the communication strategies items among the three score groups (N = 36)

Cognitive strategies, the second category, are defined as manipulation of the target language to understand and produce language, and involve three types: 1) anticipating the content, 2) using imagery to understand, think, or remember information, and 3) using notes to organize or remember information. The mean rating of the anticipating the content strategy (the red line in Figure 12) was below that of the overall (the blue line) across all assigned score groups, and there was a relatively small difference in the mean ratings among the score groups.

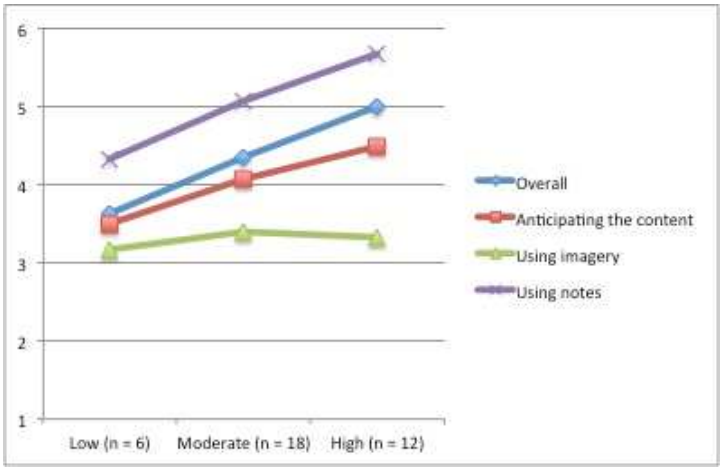


Figure 12. The ILST mean ratings of interactional authenticity on the cognitive strategies items among the three score groups ($N = 36$)

The second type of the cognitive strategies, using imagery, showed similar ratings across the assigned score groups, all around a rating of three (see the green line in Figure 12). This strategy type was perceived as moderately used, and its degree of difference was relatively small across the score groups. On the other hand, the mean rating of the third strategy type, using notes (the purple line), was higher than, but showed a similar upward pattern to that of the overall across the score groups (the blue). As in the case of reviewing and referring to notes in the communication strategies, students in the ILST group also perceived the strategy of using notes as frequently used.

Third, the metacognitive strategies, consisting of organizing, planning, and evaluating, include the following three category types: 1) setting goals for completing a given task, 2) planning the parts, sequence, or main ideas to be expressed orally, and 3) evaluating language production after task completion. Students in the high score group perceived the strategy of setting goals as used very much ($M = 5.92$), and showed a sharp increase from the mean ratings of the low and moderate score group (refer to the red line in Figure 13).

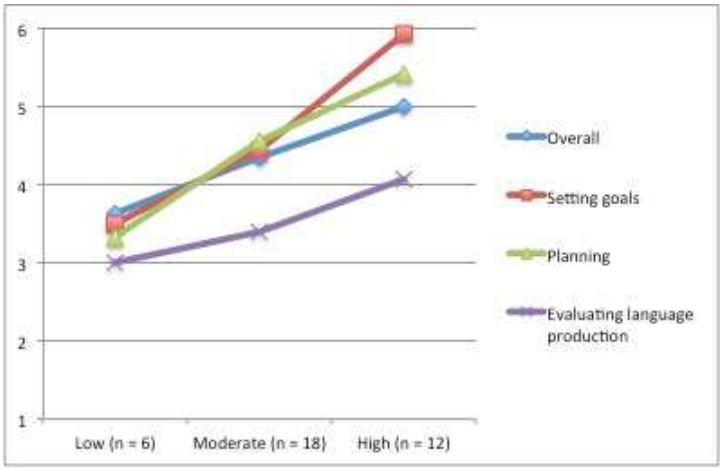


Figure 13. The ILST mean ratings of interactional authenticity on the metacognitive strategies items among the three score groups ($N = 36$)

A similar pattern was found for the strategy of planning, a great increase from the mean rating of the low score group to the high (see the green line). It appears that the first two strategy types of the metacognitive strategies, setting goals and planning, vary considerably among the three assigned score groups. According to student perception, students may have a better command of these strategy types as their academic spoken English proficiency develops. On the other hand, the strategy of evaluating language production did not increase sharply across the score groups (the purple line), and its mean ratings were lower than those of the overall.

In summary, students in the ILST group generally perceived an increased level of strategy use across the three assigned score groups, except for the strategy of using imagery that showed little group difference. Specifically, the use of two of the metacognitive strategies, setting goals and planning, increased greatly as the task scores became higher. These strategy types may be positively correlated with student English proficiency. In addition, these students reported a frequent use of strategies related to note taking such as using, reviewing, and referring to notes. It seems that note taking is an important skill for

understanding and producing language and ultimately, reaching a communicative goal of the given assessment task.

4.2.2. Multimedia-Mediated Speaking Task (MMST)

Thirty-seven students[11] among the 47 who performed the multimedia-mediated speaking task (MMST) responded to the fourteen items of the interactional authenticity questionnaire on a six-point Likert scale. Each rating indicates the extent of involvement of students' language ability, consisting of language knowledge and strategic competence, in accomplishing the MMST (1 = not at all; 6 = a lot). The following two parts present the students' perceptions of first, their degree of language knowledge and second, their degree of strategic competence, used during the MMST performance.

4.2.2.1. Language knowledge

Items 1 to 6 in the interactional authenticity questionnaire (Appendix H) concern the types of language knowledge defined in the task construct. Among the six, items 2, 3, and 4 regard the linguistic aspects related to topical differences in the input lectures (i.e., biology vs. business lecture), and accordingly, are excluded from the current analysis[12]. Only items 1, 5, and 6, which represent the core linguistic features, were considered as in the ILST analyses (see Section 4.2.1.1.). Table 20 below presents the item-level descriptive statistics of the three ratings from the group of students assigned to the MMST.

[11] As explained in Footnote 9, only the students of the Fall 2014 administration completed the questionnaire, therefore, 37 students.
[12] The analysis of the ratings on these items will be reported in Section 4.3.2.

Table 20. Descriptive statistics for the MMST item-level ratings of interactional authenticity: Language knowledge ($N = 37$)

	Interactional Authenticity Questionnaire Items: Language Knowledge	Mean	Median	Mode	SD	Min	Max
1	Relational processes: Identifying	4.30	4.00	4	1.18	2	6
5	Existential processes	3.78	4.00	4	1.65	1	6
6	Pronouns: Exophoric reference	2.95	3.00	2	1.53	1	6

Different from the varying degrees of the ILST student perception reported in Section 4.2.1.1., the three item-level mean ratings from the MMST group all lie in the moderate degree of involvement, ranging from a mean of 2.95 (item 6) to 4.30 (item 1). First, students perceived knowledge of identifying relational processes (item 1) only moderately used during their task performance ($M = 4.30$), compared to the highly perceived degree of involvement by students in the ILST group ($M = 4.78$). For example, Student 2.16, one of the 11 students who gave item 1 a rating of four said, "[i]t's just to give the narrow and the broad," which means he just needed to define two aspects, a narrow and a broad definition of a tool, so the degree of involved knowledge of identifying relational processes was moderate. Interestingly, students in the two groups, the ILST and the MMST, perceived using different levels of knowledge of identifying relational processes under the same task feature of defining two entities. Second, knowledge of existential processes was perceived by students in the MMST group as moderately involved ($M = 3.78$) as by those in the ILST group ($M = 3.83$).

Third, students in the MMST group also responded that they used moderate levels of knowledge of pronouns as exophoric reference ($M = 2.95$). This differed from responses by students in the ILST group who reported low involvement of this linguistic aspect ($M = 2.58$). More contextual information in the MMST input lectures may have contributed to a

slightly better simulation of the imagined audience and tapped to a greater extent the student knowledge of pronouns as exophoric reference.

To examine the degree of interactional authenticity across three sub-groups of students according to their assigned task scores (low, moderate or high), the item-level mean ratings and standard deviations were calculated (refer to Table 21).

Table 21. Descriptive statistics for the MMST item-level ratings of interactional authenticity per assigned task score: Language knowledge ($N = 37$)

	Interactional Authenticity Questionnaire Items: Language Knowledge	Mean (Standard Deviation)		
		Low ($n = 8$)	Moderate ($n = 21$)	High ($n = 8$)
1	Relational processes: Identifying	3.75 (1.28)	4.33 (1.02)	4.75 (1.39)
5	Existential processes	3.13 (1.64)	3.86 (1.77)	4.25 (1.28)
6	Pronouns: Exophoric reference	4.00 (1.77)	2.62 (1.43)	2.75 (1.17)

Students in these three sub-groups perceived the extent of interactional authenticity on the first two items similarly. First, students perceived similar degrees regarding how much knowledge of identifying relational processes (item 1) was used during their MMST performance, with a slightly increasing trend (refer to the blue line in Figure 14).

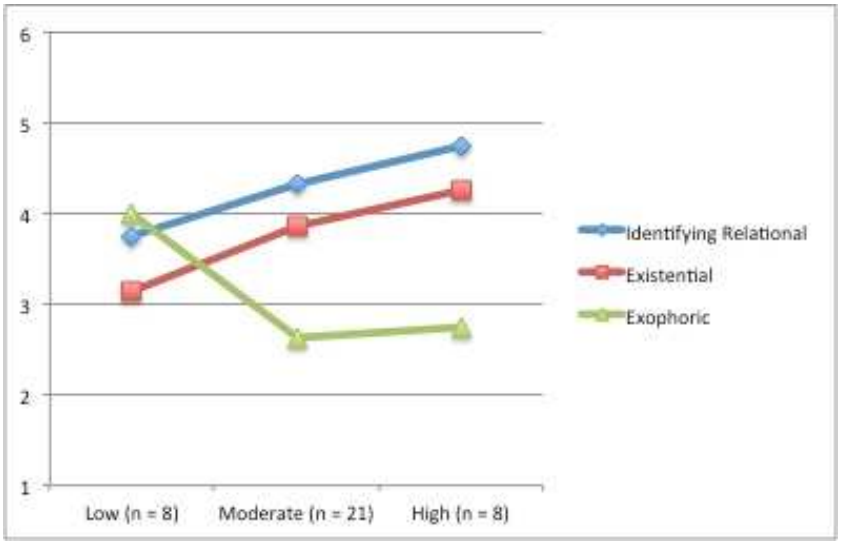


Figure 14. The MMST mean ratings of interactional authenticity on the language knowledge items among the three score groups ($N = 37$)

Second, students also perceived similar levels of using knowledge of existential processes (item 5) with a slightly increasing trend (see the red line in Figure 14). On the other hand, interestingly, eight students in the low score group showed a higher degree of involved knowledge of pronoun used as exophoric reference ($M = 4.00$) than did those in the moderate score group ($M = 2.62$) and the high score group ($M = 2.75$), as visually shown in the green line of Figure 14.

4.2.2.2. Strategic competence

Items 7 to 14 in the interactional authenticity questionnaire (Appendix H) concern the types of strategic competence defined in the task construct. Table 22 below presents the item-level descriptive statistics of the ratings from the students who were assigned to the MMST. The mean of the item statistics denotes the average rating across all eight questionnaire items, considered as the overall degree of perceived interactional authenticity of the MMST regarding strategic competence. This figure allows an interpretation on the same scale as the item statistics.

Table 22. Descriptive statistics for the MMST item-level ratings of interactional authenticity: Strategic competence ($N = 37$)

	Interactional Authenticity Questionnaire Items: Strategic competence	Mean	Median	Mode	SD	Min	Max
7	Setting goals	4.41	5.00	5	1.36	1	6
8	Anticipating the content	4.54	5.00	5	1.48	1	6
9	Using imagery	4.24	5.00	5	1.62	1	6
10	Using notes	5.03	5.00	6	1.14	2	6
11	Reviewing notes	4.57	5.00	6	1.43	2	6
12	Planning	4.43	5.00	5	1.43	1	6
13	Referring to notes	4.73	5.00	5	1.17	1	6
14	Evaluating language production	3.68	4.00	2	1.51	1	6
	Overall	4.45	4.75	4.50	0.94	1.50	5.88

As in the analysis of the ILST data, the mean of the item-level mean ratings was treated as indicating the students' overall perception of interactional authenticity regarding strategic competence. The scale of interactional authenticity (strategic competence) of the MMST had high reliability, Cronbach's α = .83. The overall perception was moderate (M = 4.45, SD = 0.94). Students in both the ILST and the MMST group had similar overall perceptions (M = 4.44, SD = 0.90 for the ILST; M = 4.45, SD = 0.94 for the MMST).

The item-level mean ratings range from a moderate (e.g., M = 3.68 for item 14) to high degree of involvement (e.g., M = 5.03 for item 10). As among students in the ILST group, students assigned to the MMST reported a high degree of involvement regarding strategies related to note taking: 1) using notes to organize or remember information while listening to the input lecture (M = 5.03, item 10), 2) reviewing notes to remember or formulate what to say during response preparation (M = 4.57, item 11), and 3) referring to notes to remember or formulate what to say while speaking (M = 5.00, item 13). Notes seemed to be an important element in task completion for students in the MMST group as well.

Noticeably, the strategy of using imagery to understand, think, or remember information (item 9) was perceived as used more frequently by students in the MMST group (M = 4.24) than by those in the ILST group (M = 3.33). The content visuals included in the MMST input lectures may have facilitated students' understanding and recall of the information, and led to their reporting a frequent use of imagery. For example, Student 1.5, one of the 21 students who assigned item 9 a rating of five or six, reported the content visuals helped him to understand the lecture: "… he [the professor in the lecture] used image, and this and this thing describe what he say and make us understand it easily." Student 3.6,

another student in the high perception group, pointed out the usefulness of the content visuals in remembering the information, as reported in the following quote: "… one visual aids I remember was he gave a example, he had it because of class, so that makes it easier to remember certain information and how he relates the terms to the pictures that he showed me."

To examine the degree of interactional authenticity regarding strategic competence across three sub-groups according to the assigned task scores (low, moderate, or high), the item-level mean ratings and standard deviations per score group were calculated (refer to Table 23). As presented in sub-section 4.2.1.2, the eight aspects of strategic competence in the interactional authenticity questionnaire are categorized and labeled as one of the three types of strategies derived from Swain et al., (2009): 1) communication strategies (items 11 and 13), 2) cognitive strategies (items 8, 9, 10), and 3) metacognitive strategies (items 7, 12, and 14).

Table 23. Descriptive statistics for the MMST item-level ratings of interactional authenticity per assigned task score: Strategic competence ($N = 37$)

	Interactional Authenticity Questionnaire Items: Strategic Competence	Mean (Standard Deviation)		
		Low ($n = 8$)	Moderate ($n = 21$)	High ($n = 8$)
	Communication Strategies			
11	Reviewing notes	4.00 (1.85)	4.57 (1.40)	5.13 (0.84)
13	Referring to notes	4.38 (1.69)	4.71 (1.10)	5.13 (0.64)
	Cognitive Strategies			
8	Anticipating the content	3.88 (2.03)	4.48 (1.37)	5.38 (0.74)
9	Using imagery	4.13 (2.03)	4.29 (1.62)	4.25 (1.39)
10	Using notes	4.50 (1.41)	5.10 (1.14)	5.38 (0.74)
	Metacognitive Strategies			
7	Setting goals	4.50 (1.31)	4.14 (1.53)	5.00 (0.76)
12	Planning	4.00 (2.07)	4.38 (1.20)	5.00 (1.20)
14	Evaluating language production	4.13 (1.89)	3.38 (1.43)	4.00 (1.31)
	Overall	4.19 (1.35)	4.38 (0.87)	4.91 (0.49)

The overall degree of perceived interactional authenticity regarding strategic competence showed an increasing trend across the three assigned task scores (M = 4.19 for low; M = 4.38 for moderate; M = 4.91 for high), depicted as the blue line in the next three figures (Figures 15, 16, and 17), but was less steep than that of the student perception in the ILST group (refer to sub-section 4.2.1.2.). This tendency across the three scores varies slightly per sub-strategy.

The mean ratings of the strategy of reviewing notes (the red line in Figure 15), the first strategy type under the category of communication strategies, were similar to those of the overall (the blue line) across all three assigned task scores, although the increase was relatively sharp compared to that of the overall. On the other hand, the mean ratings for the strategy of referring to notes across three score groups (the green line) increased to a similar degree as those of the overall, but were slightly higher throughout.

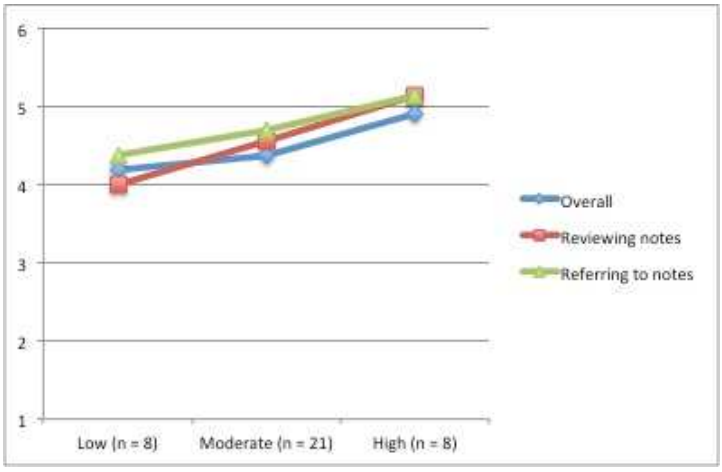


Figure 15. The MMST mean ratings of interactional authenticity on the communication strategies items among the three score groups (N = 37)

Cognitive strategies, the second sub-strategy, involve three types, 1) anticipating the content, 2) using imagery to understand, think, or remember information, and 3) using notes to organize or remember information. The mean ratings for the strategy of anticipating the

content across the three assigned task scores (refer to the red line in Figure 16) increased sharply relative to those of the overall (the blue line). However, the mean ratings for the strategy of using imagery (the green line) showed little difference across the score groups. The degree of perceived use for the third type, using notes (the purple line), increased to a similar extent as that of the overall, but the mean ratings were higher throughout the task score groups.

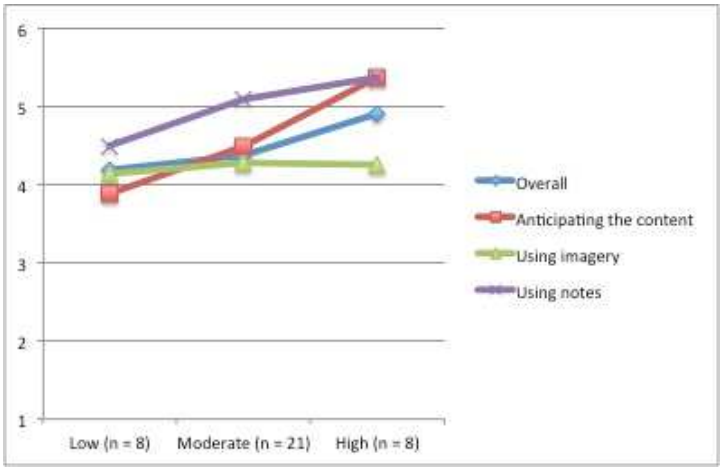


Figure 16. The MMST mean ratings of interactional authenticity on the cognitive strategies items among the three score groups ($N = 37$)

The mean ratings of the strategy of setting goals (the red line in Figure 17), the first strategy type under the category of metacognitive strategies, were similar to those of the overall (the blue line) across all three assigned task scores. In addition, students in the MMST group perceived their use of the planning strategy (the green line) as similar to that of the overall, but with a slightly sharper increase across the assigned task scores. The last strategy under the category of metacognitive strategies, evaluating language production (the purple line), showed an interesting pattern. Different from the other types of strategies, the perceived level of this strategy type generally decreased across the task score groups.

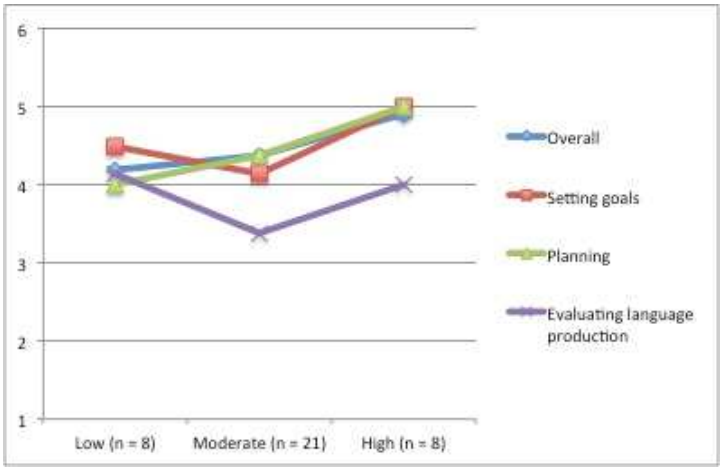


Figure 17. The MMST mean ratings of interactional authenticity on the metacognitive strategies items among the three score groups ($N = 37$)

In summary, students in the MMST group generally perceived an increased level of strategy use across the three assigned score groups, except for the strategies of using imagery that showed little group difference and evaluating language production that showed a slight decrease. Specifically, use of the strategy of anticipating the content increased greatly as the task scores became higher. This strategy type may be positively correlated with student English proficiency. In addition, as indicated by students in the ILST group, students in the MMST group also reported a frequent use of the strategy of using notes. For this assessment task, note taking seems an important component in achieving the desired communicative goal.

4.2.3. Comparison Between the ILST and the MMST

Sub-sections 4.2.1 and 4.2.2 discussed the perceived interactional authenticity of the integrated listening-speaking task (ILST) and the multimedia-mediated speaking task (MMST), respectively, and also across the three assigned task scores. In this sub-section, comparisons between the two assessment tasks on the perceived interactional authenticity

regarding first, language knowledge and second, strategic competence, are made based on the results of the independent *t*-tests for language knowledge and one-way multivariate analysis of variance (MANOVA) for strategic competence. The visual representations of line graphs and the findings of two-way analysis of variance (ANOVA) are used for comparisons per assigned task score.

4.2.3.1. Language knowledge

The ratings on the three selected interactional questionnaire items regarding language knowledge, namely items 1, 5, and 6, from 36 students in the ILST group and 37 students in the MMST group were statistically compared (73 students in total), using independent *t*-tests. To control the familywise error rate from three statistical tests conducted on the same data and ensure that the cumulative Type I error remains below .05, I used .017 ($P_{Crit} = .05/3$) as my criterion for significance, following the Bonferroni correction. First, on average, students in the ILST group perceived more frequent use of identifying relational processes ($M = 4.78$, $SE = 0.22$) than did those in the MMST group ($M = 4.30$, $SE = 0.19$). However, this difference, 0.48, BCa 95% CI [- 0.093, 1.032], was not statistically significant $t(71) = 1.65$, $p = .104$.

Second, on average, students in both the ILST ($M = 3.83$, $SE = 0.29$) and the MMST group ($M = 3.78$, $SE = 0.27$) perceived their use of knowledge of existential processes similarly, and the mean difference, 0.05, BCa 95% CI [- 0.684, 0.817], was also not statistically significant $t(71) = 0.12$, $p = .901$. On the other hand, on average, students in the ILST group ($M = 2.58$, $SE = 0.30$) perceived less frequent use of pronouns as exophoric

reference than did those in the MMST group (M = 2.95, SE = 0.25). However, the difference, - 0.36, BCa 95% CI [- 1.160, 0.414], was not statistically significant $t(71) = - 0.93$, $p = .358$.

Overall, although there seemed to be mean differences between the ILST and the MMST group based on the item statistics, a series of independent *t*-tests showed that students in the two groups considered the degree of the interactional authenticity regarding language knowledge of both assessment tasks as similar for each authenticity aspect. According to student perception, both types of integrated assessment tasks do not function differently in tapping students' knowledge of the target grammatical features defined in the task construct and, in other words, do not differ in eliciting students' language knowledge.

In addition, comparisons per assigned task score also did not reveal major differences in student perception of interactional authenticity regarding language knowledge between the ILST and the MMST group. First, although students in both the ILST and the MMST group perceived increased use of knowledge of identifying relational processes across the three assigned scores (low, moderate, and high) (see Figure 18), the degree of perception increased to a greater extent among students in the ILST group.

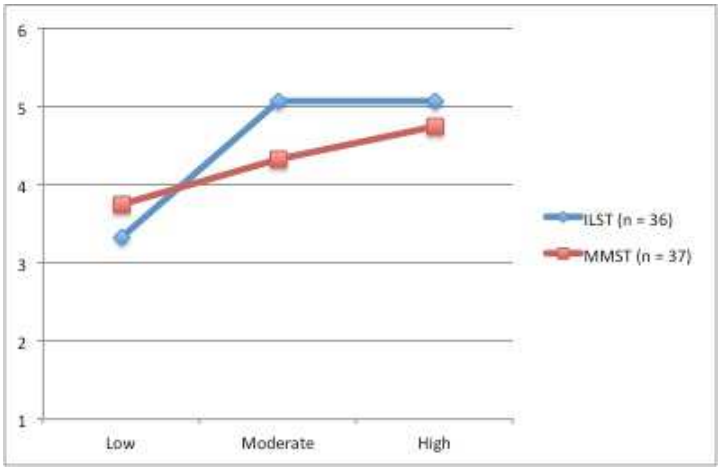


Figure 18. The ILST and MMST mean ratings of interactional authenticity on knowledge of identifying relational processes among the three score groups (N = 73)

Specifically, the mean rating of the ILST students in the moderate score group ($M = 5.06$) showed a sharp increase from those of the low score group ($M = 3.33$), and there was a relatively large difference in the mean ratings between the ILST and the MMST-moderate score group. However, a two-way ANOVA revealed that there was a non-significant interaction between the task type and the assigned task score, on the perceived level of using knowledge of identifying relational processes, $F(2, 67) = 1.22$, $p = .301$; in other words, students in the ILST and the MMST group did not significantly differ in their perception across the assigned scores.

Second, students in both the ILST and the MMST group also perceived an increased use of knowledge of existential processes across the three assigned scores (see Figure 19), although the degree of perception increased to a greater extent among students in the ILST group. In particular, the mean rating of the ILST students in the low score group ($M = 2.33$) appeared to have a relatively large difference from that of the MMST low score group ($M = 3.13$).

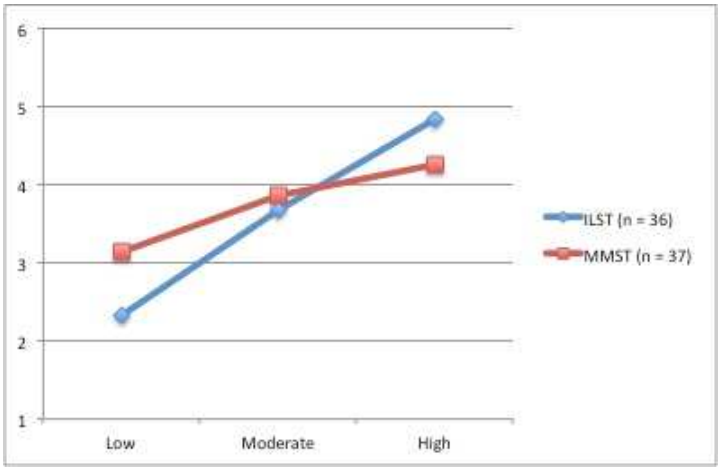


Figure 19. The ILST and MMST mean ratings of interactional authenticity on knowledge of existential processes among the three score groups ($N = 73$)

However, a two-way ANOVA revealed that there was a non-significant interaction between the task type and the assigned task score, on the perceived level of using knowledge of existential processes, $F(2, 67) = 0.76$, $p = .470$; in other words, students in the ILST and the MMST group did not significantly differ in their perception across the assigned scores.

Third, students in the two groups differed in their perception of using knowledge of pronouns as exophoric reference across the three assigned scores, with the ILST group perceiving increased use but the MMST group perceiving decreased use (see Figure 20). Particularly, the mean rating of the MMST students in the low score group ($M = 4.00$) was substantially different from that of the ILST low score group ($M = 1.83$).

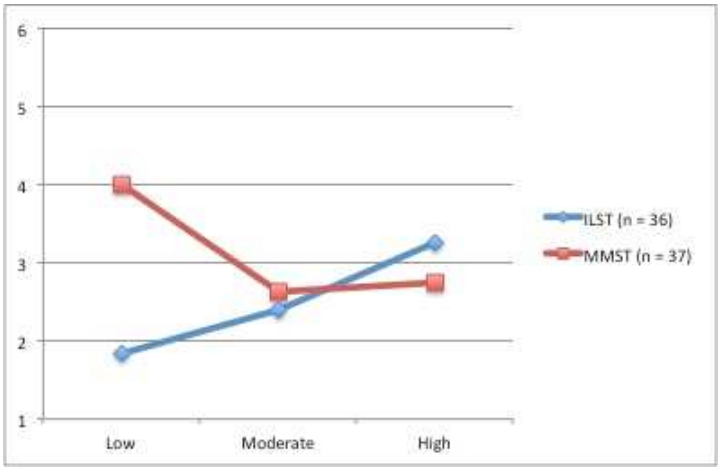


Figure 20. The ILST and MMST mean ratings of interactional authenticity on knowledge of pronouns used as exophoric reference among the three score groups ($N = 73$)

However, a two-way ANOVA revealed that there was a non-significant interaction between the task type and the assigned task score, on the perceived level of using knowledge of pronouns as exophoric reference, $F(2, 67) = 2.80$, $p = .068$; in other words, students in the ILST and the MMST group did not significantly differ in their perception across the assigned scores.

To sum up, the perception of students in both the ILST and the MMST group on the degree of using the three aspects of language knowledge did not significantly differ. These non-significant group differences were also identified across the assigned task scores. As in the findings of the perceived situational authenticity (refer to sub-section 4.1), students in both groups did not differ in their perception of their language knowledge use during their task performance; in other words, the perceived interactional authenticity regarding language knowledge was the same for both groups.

4.2.3.2. Strategic competence

The ratings on the eight interactional questionnaire items regarding strategic competence (items 7 to 14) from 36 students in the ILST group and 37 students in the MMST group were statistically compared (73 students in total), using one-way multivariate analysis of variance (MANOVA). Using Wilk's lambda, there was not a significant difference between the ILST and the MMST in the ratings on the eight interactional questionnaire items regarding strategic competence collectively, $\Lambda = 0.84$, $F(8, 64) = 1.32$, $p = .174$.

Separate univariate ANOVAs on the questionnaire ratings also revealed non-significant group differences on each aspect of the perceived interactional authenticity with one exception: 9) using imagery, $F(1, 71) = 4.86$, $p = .031$. The ones that showed non-significant differences were: 7) setting goals, $F(1, 71) = 1.44$, $p = .235$, 8) anticipating the content, $F(1, 71) = 1.47$, $p = .230$, , 10) using notes, $F(1, 71) = 0.15$, $p = .696$, 11) reviewing notes, $F(1, 71) = 1.32$, $p = .254$, 12) planning, $F(1, 71) = 0.39$, $p = .537$, 13) referring to notes, $F(1, 71) = 1.08$, $p = .301$, and 14) evaluating language production, $F(1, 71) = 0.13$, $p = .715$. Students in the two groups, the ILST and the MMST, considered the degree of the

interactional authenticity of using imagery different, $F(1, 71) = 4.86$, $p = .031$, unlike that of the other authenticity aspects perceived as similar between the groups. The content visuals included in the MMST input lectures appeared to facilitate a frequent use of imagery in understanding and remembering of the information for accomplishing the assigned task.

Comparisons per assigned task score did not reveal major differences in student perception of interactional authenticity regarding strategic competence between the ILST and the MMST group. First, students' overall perception of interactional authenticity regarding strategic competence was not significantly different between the ILST and the MMST group, $F(2, 67) = 0.61$, $p = .546$. The mean rating of the ILST-low group (the first point of the blue line in Figure 21) was slightly lower than that of the MMST low group (on the red line) and therefore the mean ratings in the ILST group gradually increased across the assigned task scores. However, this difference was not statistically significant. Students in both groups across the assigned task scores perceived a similar degree of interactional authenticity regarding strategic competence overall.

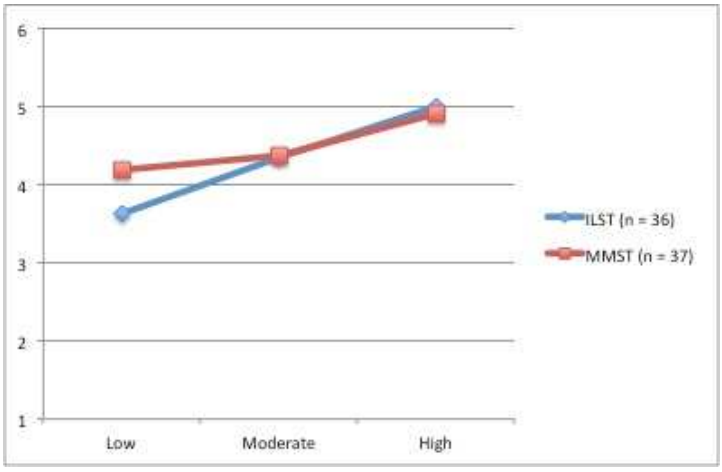


Figure 21. The ILST and MMST mean ratings of interactional authenticity regarding strategic competence on the overall among the three score groups ($N = 73$)

Two-way ANOVAs on the eight aspects of interactional authenticity regarding strategic competence also showed non-significant interactions between the task types (the ILST or the MMST) and the assigned task scores (low, moderate, and high) on the student perception of using the following strategy types: 7) setting goals, $F(2, 67) = 2.69$, $p = .075$, 8) anticipating the content, $F(2, 67) = 0.17$, $p = .842$, 9) using imagery, $F(2, 67) = 0.00$, $p = .999$, 10) using notes, $F(2, 67) = 0.18$, $p = .836$, 11) reviewing notes, $F(2, 67) = 0.30$, $p = .742$, 12) planning, $F(2, 67) = 0.70$, $p = .500$, 13) referring to notes, $F(2, 67) = 0.25$, $p = .781$, and 14) evaluating language production, $F(2, 67) = 0.97$, $p = .384$. Students in both the ILST and the MMST group across the assigned task scores perceived a similar degree of interactional authenticity regarding all eight aspects of strategic competence.

4.2.4. Section Summary

To address the second research question, the ratings of *interactional authenticity* regarding language knowledge and strategic competence assigned by students who took either of the two assessment tasks, the ILST and the MMST, were analyzed. For language knowledge, students in the ILST group perceived a relatively frequent use of their knowledge of identifying relational processes, one of the types of target language knowledge defined in the task construct ($M = 4.78$, $SD = 1.31$) on a six-point Likert scale. On the other hand, they perceived a relatively moderate use of their knowledge of the other two aspects of the target language knowledge, 1) knowledge of existential processes and 2) knowledge of pronouns as exophoric reference ($M = 3.83$, $SD = 1.75$ for existential processes; $M = 2.58$, $SD = 1.81$ for pronouns as exophoric reference). Perceived degrees of using all three aspects of the target language knowledge increased across the assigned task scores.

On the other hand, students in the MMST group perceived a relatively moderate use of their knowledge of identifying relational processes ($M = 4.30$, $SD = 1.18$). However, this group difference was not significant $t(71) = 1.65$, $p = .104$. Students in the MMST group also perceived a relatively moderate use of their knowledge of the other two aspects of the target language knowledge, 1) existential processes and 2) pronouns as exophoric reference ($M = 3.78$, $SD = 1.65$ for existential processes; $M = 2.95$, $SD = 1.53$ for pronouns as exophoric reference). These also do not significantly differ from those for the ILST group, $t(71) = 0.12$, $p = .901$ for knowledge of existential processes and $t(71) = - 0.93$, $p = .358$ for knowledge of pronouns as exophoric reference. Whereas perceived degrees of using knowledge of identifying relational and existential processes increased across the assigned task scores, those of using knowledge of pronouns as exophoric reference seemingly decreased. However, two-way ANOVAs revealed that there were non-significant interactions between the task type and the assigned task score on the perceived level of using all aspects of language knowledge, $F(2, 67) = 1.22$, $p = .301$ for identifying relational processes, $F(2, 67) = 0.76$, $p = .470$ for existential processes, and $F(2, 67) = 2.80$, $p = .068$ for pronouns as exophoric reference.

For strategic competence, students in the ILST group perceived a relatively moderate degree of using their strategic competence overall ($M = 4.44$, $SD = 0.90$) on a six-point Likert scale. The item-level mean ratings range from a moderate (e.g., $M = 3.33$ for the strategy of using imagery) to a high degree of involvement (e.g., $M = 5.14$ for the strategy of using notes). Perceived degrees of all aspects of the target strategic competence increased across the assigned task scores.

Students in the MMST group also perceived a relatively moderate degree of using their strategic competence overall (M = 4.45, SD = 0.94). This also did not significantly differ from that of the ILST group, Λ = 0.84, $F(8, 64) = 1.32$, p = .174. The item-level mean ratings range from a moderate (e.g., M = 3.68 for the strategy of evaluating language production) to a high degree of involvement (e.g., M = 5.03 for the strategy of using notes). Separate univariate ANOVAs also revealed non-significant group differences on each aspect of the perceived interactional authenticity, except for the strategy of using imagery, $F(1, 71) = 4.86$, p = .031. Across the assigned task scores, perceived degrees of all aspects of interactional authenticity regarding strategic competence increased, except those of using imagery and evaluating language production, which changed little across the task scores. However, two-way ANOVAs revealed that there were non-significant interactions between the task type and the assigned task score on the perceived level of using all aspects of strategic competence.

To sum up, students perceived a similar degree of interactional authenticity regarding language knowledge elicited by the two assessment tasks, the ILST and the MMST. Similar degrees of perception of both assessment tasks were also found across the assigned task scores. Students perceived a similar degree of interactional authenticity regarding strategic competence of the two assessment tasks, except for the strategy of using imagery. It seems that content visuals included in the MMST facilitated a more frequent use of imagery to understand and remember the contents of the lecture, and therefore a higher perception of using this strategy type indicated by students in the MMST group. However, students in both the ILST and the MMST group perceived similar use of all aspects of interactional authenticity regarding strategic competence across the assigned task scores. In general, both

the ILST and the MMST functioned similarly in eliciting students' language ability consisting of language knowledge and strategic competence.

4.3. Elicited Language Knowledge Manifested in Task Performance

This section reports on the results of the third research question: What language knowledge do the two assessment tasks—the integrated listening-speaking task (ILST) and the multimedia-mediated speaking task (MMST)—elicit in the examinees' spoken response? How different is this elicited language knowledge between the tasks and across assigned task scores? The data consist of pruned transcripts of the 93 one-minute examinee spoken responses to either the ILST or the MMST. Counts of several linguistic features identified in each of the spoken responses were used to interpret the types and extent of elicited language knowledge of the examinees in the two assessment tasks.

In the first sub-section, I will present the descriptive statistics for processes, one of the features that refers to the meaning of the verbs, used in student spoken responses to the ILST and the MMST, and statistically compare the use of the two key processes, defined in the task construct, between the two assessment tasks and also across the assigned three task scores: low, moderate and high. The second sub-section discusses the use of processes between two versions, biology and business, of the ILST and the MMST. I will compare the use of processes between the input lecture and student spoken responses, illustrated with a few quotes from student responses that were particularly revealing of the findings. In the third sub-section, I will provide the descriptive statistics for pronouns used as exophoric reference (something outside of the text itself) such as "we" and "you" in student spoken

responses, and statistically compare their use between the two assessment tasks and across the assigned task scores. The last sub-section will present a summary.

4.3.1. Use of Processes for Task Types: The ILST and the MMST

This sub-section addresses the first sub-research question three regarding the use of processes, in other words, the meaning of the verbs in clauses, between the two task types (the ILST and the MMST), and also across the three assigned task scores (low, moderate, and high). Forty-six students who took the ILST and 47 who took the MMST provided a one-minute spoken response to their assigned task. These spoken responses were transcribed, pruned, and identified as being one of the seven types of processes in the transitivity system, (refer to 3.6.3.1.). Each process type found in each spoken response was counted, and these counts were compared across the groups. The first part of this sub-section presents comparisons of the two target processes, identifying relational and existential processes, used in the ILST and the MMST spoken responses. Then, the same kinds of comparisons across assigned task scores follow in the second sub-section. The final sub-section provides a summary.

4.3.1.1. Comparisons between the ILST and the MMST

To compare the use of processes among students who took the ILST and the MMST, descriptive statistics for the seven types of processes used in student spoken responses were computed, and are presented in Table 24.

Table 24. Descriptive statistics for processes used in student spoken responses to the ILST and the MMST ($N = 93$)

Process	ILST ($n = 46$)			MMST ($n = 47$)		
	Mean	Median	Range	Mean	Median	Range
Material	2.54	2.00	0 – 7	2.38	2.00	0 – 7
Mental	0.72	0.00	0 – 6	0.47	0.00	0 – 5
Verbal	0.33	0.00	0 – 3	0.23	0.00	0 – 3
Behavioral	0.41	0.00	0 – 3	0.36	0.00	0 – 2
Relational: Identifying	2.96	3.00	0 – 6	3.38	3.00	0 – 8
Relational: Attributive	0.63	0.00	0 – 3	0.72	0.00	0 – 4
Existential	0.39	0.00	0 – 2	0.64	1.00	0 – 3
Total	7.96	8.00	1 – 14	8.19	8.00	3 – 17

As expected, identifying relational processes were used most by students in both groups ($M = 2.96$, $Mdn = 3.00$ for the IMST; $M = 3.38$, $Mdn = 3.00$ for the MMST), as this is one of the key process types in the assessment tasks that require explaining two aspects of a concept. For example, a successful spoken response that explains two factors of product quality, reliability and features, is expected to include identifying relational processes, "to be," such as "the first one **is** reliability, and the second one **is** the features of a product" (identifying relational processes highlighted in bold). This finding matches the ILST student perception of high use of this process in their spoken responses (see Section 4.2.1.1.). Although the MMST students perceived a moderate use of this process (see Section 4.2.2.1.), they, in fact, used it frequently as did the ILST students.

Existential processes (e.g., "there are") were not used as often by the ILST students ($M = 0.39$, $Mdn = 0.00$), whereas the MMST students used this process type moderately in their spoken responses ($M = 0.64$, $Mdn = 1.00$). Although students in both groups perceived a moderate use of existential processes (see Sections 4.2.1.1. and 4.2.2.1.), only the MMST students' perceptive and actual use of this process matched.

To statistically compare the use of these two target processes, identifying relational and existential, between the two assessment tasks, contingency tables for these processes were created and Pearson's chi-square tests were conducted. Table 25 below is a contingency table presenting the number of students who used identifying relational processes less than twice, or twice or more, depending on their assigned assessment task. Two is a cut-off point as it is expected that students need to use identifying relational processes at least twice to explain two aspects of an academic concept.

Table 25. Contingency table showing the number of students who used identifying relational processes depending on the assigned assessment task (N = 93)

		Task		
		ILST	MMST	Total
Identifying Relational Process	< twice	9 (19.6%)	6 (12.8%)	15 (16.1%)
	≥ twice	37 (80.4%)	41 (87.2%)	78 (83.9%)
	Total	46 (100.0%)	47 (100.0%)	93 (100.0%)

More than 80% of the students in both groups (ILST: 80.4%; MMST: 87.2%) used identifying relational processes (e.g., "to be") twice or more. This indicates that for around 80% of students, both assessment tasks tapped student knowledge of identifying relational process and the groups did not differ significantly regarding the association between the type of assessment tasks and the use of identifying relational processes in student spoken response $\chi^2(1) = 0.79, p < .373$.

A contingency table presenting the number of students who used existential processes (e.g., "there are") depending on their assigned assessment task is presented in Table 26. More than half of the MMST students (55.3%) used existential processes to classify two aspects of an academic concept, although only approximately a third of the ILST students (32.6%) used

this process type. A successful spoken response that introduces two factors of product quality, reliability and features, is expected to include a relational process, "there are," such as "**there are** two factors of product quality" (existential processes highlighted in bold).

Table 26. Contingency table showing the number of students who used existential processes depending on the assigned assessment task ($N = 93$)

		Task		
		ILST	MMST	Total
Existential Process	Yes	15 (32.6%)	26 (55.3%)	41 (44.1%)
	No	31 (67.4%)	21 (44.7%)	52 (55.9%)
	Total	46 (100.0%)	47 (100.0%)	93 (100.0%)

There was a statistically significant association between the type of assessment tasks and whether or not students used existential processes in their spoken response $\chi^2(1) = 4.86$, $p < .027$. Based on the odds ratio, the odds of students using existential processes were 2.56 times higher if they took the MMST than the ILST. It appears that the MMST functioned better in eliciting the use of existential processes and this helped with classification of a concept that needs to be explained. The MMST possibly allowed students to visualize the classification and include such information in their oral explanation, using an existential process.

4.3.1.2. Comparisons between the ILST and the MMST per task score

It is expected that students who have a higher English proficiency and receive a higher score on the assigned assessment task will have more involvement in using the targeted language knowledge defined in the task construct. To statistically compare the use of the two target processes, identifying relational and existential, between the two assessment

tasks and also across the assigned task scores (low, moderate, and high), contingency tables for the two processes were created and three-way loglinear analyses were conducted. First, a contingency table presenting the number of students who used identifying relational processes less than twice or twice or more depending on their assigned assessment task and task score is presented in Table 27 below. Two is a cut-off point again as students are expected to use identifying relational processes at least twice for defining two aspects of an academic concept.

Table 27. Contingency table showing the number of students who used identifying relational processes depending on the assigned assessment task and task score ($N = 93$)

		Task						Total
		ILST			MMST			
		Low	Moderate	High	Low	Moderate	High	
Identifying Relational Process	< twice	3	5	1	2	3	1	15
		33%	23%	7%	18%	11%	11%	16%
	≥ twice	6	17	14	9	24	8	78
		67%	77%	93%	82%	89%	89%	84%
	Total	9	22	15	11	27	9	93
		100%	100%	100%	100%	100%	100%	100%

The ratios of the number of students who used this process twice or more to those of the number of students who used it less than twice were larger for the MMST students who received a low or moderate task score. On the other hand, the ILST students who received a high task score were more likely to use identifying relational processes twice or more relative to less than twice than did the MMST students who received a high task score.

To test for significance of the associations, a three-way loglinear analysis, the three-category version of a chi-square test, was used. This analysis produced a final model that retained only the one-way effects. The likelihood ratio of this model was $\chi^2(0) = 0$, $p = 1$. This indicated that all the two-way interactions and the three-way interaction collectively

were not significant, $\chi^2(7) = 6.46$, $p < .487$, and specifically, the three-way interaction alone was not significant, either, $\chi^2(2) = 0.73$, $p < .695$. This analysis reveals that the association between the type of assessment tasks and the use of identifying relational processes in student spoken response did not differ significantly depending on the assigned task score. Contrary to expectations, the degree of students' use of the targeted language knowledge did not significantly increase as their assigned task scores were higher.

The number of students who used existential processes, the second target process, less than twice or twice or more depending on their assigned assessment task and task score is presented in a contingency table below. The ratios of the number of students who used this process to those who did not were larger for the MMST students across all three assigned task scores. Interestingly, for the MMST students, a chance of using existential processes appears to increase as assigned task scores become higher.

Table 28. Contingency table showing the number of students who used existential processes depending on the assigned assessment task and task score ($N = 93$)

		Task						Total
		ILST			MMST			
		Low	Moderate	High	Low	Moderate	High	
Existential Process	Yes	3	8	4	5	15	6	41
		33%	36%	27%	45%	56%	67%	44%
	No	6	14	11	6	12	3	52
		67%	64%	73%	55%	44%	33%	56%
	Total	9	22	15	11	27	9	93
		100%	100%	100%	100%	100%	100%	100%

However, the three-way loglinear analysis produced a final model that retained only the one-way effects. The likelihood ratio of this model was $\chi^2(0) = 0$, $p = 1$. This indicated that all the two-way interactions and the three-way interaction collectively were not significant, $\chi^2(7) = 8.43$, $p < .296$, and specifically, the three-way interaction alone was not

significant, either, $\chi^2(2) = 1.01$, $p < .604$. This analysis reveals that the association between the type of assessment tasks and the use of existential processes in student spoken response did not differ significantly depending on the assigned task score. This is contrary to the expectation of positive correlation between students' use of the targeted language knowledge and the assigned task scores.

4.3.1.3. Summary

It was expected that identifying relational processes were frequently used, relative to other types of process, to achieve the main task goal of defining academic aspects and overall, both the ILST and the MMST students used this first target process, frequently (M = 2.96, Mdn = 3.00 for the IMST; M = 3.38, Mdn = 3.00 for the MMST) as expected. The finding is consistent with the frequent use of knowledge of this process perceived by the ILST students (see Section 4.2.1.1.), but does not match the perception of moderate use by the MMST students (see Section 4.2.2.1.).

More than 80% of the students in both groups (ILST: 80.4%; MMST: 87.2%) used identifying relational processes twice or more, and their use of this process did not significantly differ between the groups $\chi^2(1) = 0.79$, $p < .373$. Moreover, the use of the process between the two groups did not significantly differ depending on the assigned task score $\chi^2(2) = 0.73$, $p < .695$. This suggests that both assessment tasks performed well on tapping student knowledge of identifying relational process regardless of students' assigned task scores.

Although students in both groups perceived a moderate use of the second target process, existential, in their spoken responses (see Sections 4.2.1.1. and 4.2.2.1.), this process

was not used as much (M = 0.39, Mdn = 0.00) by the ILST students, relative to other types of process. Only a third of the ILST students used existential process, whereas more than half of the MMST students used it. The use of the process significantly differed depending on the two assessment tasks $\chi^2(1) = 4.86$, $p < .027$. However, the use of the process between the two groups did not significantly differ depending on the assigned task score $\chi^2(2) = 1.01$, $p < .604$. It appears that the MMST functioned better on eliciting student use of existential process than did the ILST and this tendency does not differ across the assigned task scores. The content visuals included in the MMST possibly assisted students in understanding and remembering the classification of an academic concept that they were asked to orally explain and including such information in their spoken responses using an existential process.

4.3.2. Use of Processes for Task Versions: Biology and Business

Systemic functional linguistics defines a process as one type of language resource that represents the meaning of some activity (e.g., to use, to find, etc.) or way of being (e.g., to be, there are, etc.) in a situation where a language event unfolds. Choices of this linguistic element are influenced heavily by the characteristics of the context, especially the aspect of "what." In this sense, the contents of the input lecture for the two assessment tasks, the ILST and the MMST, play an important role in task performers' choice of processes in response to the lecture. In this sub-section, the relationship between types of processes in the input materials and students' choice of processes in their spoken responses is examined (i.e., the second sub-research question three).

Half of the students from the ILST and the MMST groups (23 students each) were assigned to provide a one-minute spoken explanation of an academic concept taught in a

biology lecture. The other half (23 from the ILST group and 24 from the MMST group) were given a business lecture and asked to perform the same speaking task. The spoken responses were transcribed, pruned, and identified as being one of the seven types of processes, in other words, the meaning of the verbs in clauses. The number of each process type found in each spoken response was counted, and descriptive statistics were calculated. The first part of this sub-section presents the use of processes first in the biology input lecture and second in student spoken responses to the lecture. Then, the second part presents the use of processes in the business lecture and students' spoken responses to it. The final sub-section ends with a summary.

4.3.2.1. Biology

The biology input lecture explains the two different definitions of tools, the narrow and the broad definition, with supporting examples of how sticks used by wild chimpanzees and elephants can be defined as tools. The count of the processes used in this input lecture is presented in Table 29. The table also provides percentages of processes relative to the total (33 processes), and specific examples of each process type from the lecture.

Table 29. Total counts, percentages, and examples of processes used in the biology lecture

Process	Count	Percentage	Examples
Material	16	48.5%	To change, to shape, to use, to do, to sharpen, to pull, to chew, to trim, to find, to make, to modify, to pick, and to scratch
Mental	1	3.03%	To recognize
Verbal	5	15.2%	To say
Behavioral	2	6.06%	To fit, and to depend
Relational: Identifying	7	21.2%	To be
Relational: Attributive	1	3.03%	To be
Existential	1	3.03%	There are
Total	33	100.0%	

Material processes are the most frequently used process type (48.5%) in this biology input lecture. The following lecture excerpt demonstrates an example of this frequent use of material processes, highlighted in bold.

> "To be a tool, according to the narrow definition, the object**'s gotta be** purposefully **changed** or **shaped** by the animal, or human … Wild chimpanzees **use** sticks to dig insects out of their nests . . . but most sticks lying around **won't do** the job . . . So the sticks **have to be sharpened** …The chimp **pulls** off the leaves and **chews** the stick and **trims** it down that way … The chimp **doesn't** just **find** the stick . . . it . . . uh . . . it **makes** it in a way. But the broad definition says an object **doesn't have to be modified** to be considered a tool … For example, an elephant **will** sometimes **use** a stick to scratch its back . . . it . . . it . . . just **picks** up a stick from the ground and **scratches** its back with it . . . It **doesn't modify** the stick, it **uses** it just as it's found." (Educational Testing Service, 2010)

Since the two examples, wild chimpanzees and elephants, used in the lecture involve many physical actions such as "to sharpen" and "to scratch," a material process was frequently chosen to represent such meanings.

This specific linguistic characteristic of the input lecture probably influenced the use of processes in student spoken responses where an explanation of the definitions of tools and their corresponding examples was required. In this case, a successfully completed response should include multiple instances of material processes. Table 30 presents the descriptive statistics for the seven types of processes used in student spoken responses to the biology version of the ILST and the MMST.

Table 30. Descriptive statistics for processes used in student spoken responses to the biology version of the two assessment tasks ($N = 46$)

Process	ILST ($n = 23$)			MMST ($n = 23$)		
	Mean	Median	Range	Mean	Median	Range
Material	3.83	4.00	0 – 7	3.65	4.00	0 – 7
Mental	0.26	0.00	0 – 2	0.13	0.00	0 – 1
Verbal	0.43	0.00	0 – 3	0.30	0.00	0 – 3
Behavioral	0.35	0.00	0 – 2	0.43	0.00	0 – 2
Relational: Identifying	2.96	3.00	0 – 6	3.17	3.00	0 – 8
Relational: Attributive	0.17	0.00	0 – 2	0.26	0.00	0 – 1
Existential	0.52	0.00	0 – 2	0.57	1.00	0 – 2
Total	8.52	9.00	1 – 14	8.52	8.00	3 – 17

Students in both the ILST and the MMST group typically used material processes more frequently than other processes, $M = 3.83$ for the ILST, $M = 3.65$ for the MMST, $Mdn = 4.00$ for both, *range* = 0 – 7 for both. There was a non-significant association between students' use of material processes and their assigned task type $\chi^2(7) = 9.77$, $p < .202$. Both task types seemed to succeed in eliciting students' choices of a process appropriate for representing key information required by the task. For instance, Student 3.9, one of the 15 ILST students who used a material process four (the median) or more times, used a material process such as "to change," "to pick," and "to scratch" (highlighted in bold below) four times to explain the meaning of the definitions of tools and their examples, as shown in the following excerpt.

> "… The narrow one says that like the people or the animals **change** the forms of the objects in some way and to achieve some certain uses … And in order to eat something, not only just to find a stick, it [a wild chimpanzee] **changes** the form of the stick … Just it [an elephant] randomly **pick** up the stick and **scratch** its back." (Student 3.9, ILST)

Similarly, Student 3.31, one of the 12 MMST students who used a material process four (the median) or more times, also showed a frequent use of material processes such as "to change," "to shape," "to use," and "to find" for explaining ways of defining a tool with examples that include "doings."

> "… And it [a tool] **can be changed** or **shaped**. For example, for animal chimp, and they **use** sticks … and they **can use** this stick to catch other things. … For example, for elephants, they **will found** the tool." (Student 3.31, MMST)

As students in both groups used the examples they learned from the input lecture to explain the dentitions of a tool in their spoken responses, they chose to use material processes frequently to represent the meaning of the physical actions described in the examples.

Interestingly, this highly involved level of knowledge of material processes manifested in students' production was not consistent with their only moderately perceived level of using this process type, based on the item-level mean ratings of the interactional authenticity questionnaire on a six-point Likert-scale, $M = 3.56$, $SD = 2.23$ for the ILST group, and $M = 4.00$, $SD = 1.54$ for the MMST group. Students were more extensively involved in using their knowledge of material processes when creating their spoken responses than what they perceived.

4.3.2.2. Business

The business input lecture explains the two major factors of product quality, reliability and features, and how their role in consumer decision-making has changed. The count of the processes used in this input lecture is presented in Table 31. The table also

provides percentages of processes relative to the total 26 processes, and specific examples of each process type from the lecture.

Table 31. Total counts, percentages, and examples of processes used in the business lecture

Process	Count	Percentage	Examples
Material	3	11.5%	To get, to take, and to load
Mental	5	19.2%	To determine, to assume, and to choose
Verbal	1	3.85%	To say
Behavioral	3	11.5%	To speak, to care, and to look
Relational: Identifying	10	38.5%	To mean, and to be
Relational: Attributive	4	15.4%	To be
Existential	0	0.00%	N/A
Total	26	100.0%	

To make meanings of defining the two aspects of an academic topic, which is the common rhetorical structure for both the biology and the business input lecture, an identifying relational process was used the most (38.5%). Followed by this process, a mental process was the second most frequently used (19.2%) in the business lecture. The following lecture excerpt demonstrates an example of the frequent use of this process, highlighted in bold.

> "If a consumer has to choose between two products, what **determines** the choice? **Assume** that someone, a purchaser, **is choosing** between two products that cost the same. OK? If people have a choice between two identically priced products, which one will they **choose**? They **choose** the one they think is of higher quality, of course…" (Educational Testing Service, 2011)

Since consumer decision-making, the key example used in the lecture, involves thinking processes such as "to determine" and "to choose," a mental process was chosen to represent such meanings.

The third most frequently used process (15.4%) in this business input lecture was an attributive relational process. The following lecture excerpt demonstrates an example of the frequent use of this process, highlighted in bold.

> "Well, a product **is** reliable if it works the way we expect it to work, if it can go a reasonable amount of time without needing repairs. If a product, a car for example, doesn't work the way it should and needs repairs too soon, we say it**'s** unreliable … Today it**'s** different … So reliability **is** important …"

Since the lecture describes some features of entities such as a "product," and a "car," an attributive relational process (i.e., "to be") with descriptive language (e.g., "reliable") was selected to represent such meanings.

These specific linguistic characteristics of the input lecture probably influenced the use of processes in student spoken responses where an explanation of two factors of product quality important for consumer decision-making and their relationship was required. In this case, a successfully completed response should include multiple instances of mental and attributive relational processes. Table 32 presents the descriptive statistics for the seven types of processes used in student spoken responses to the business version of the ILST and the MMST.

Table 32. Descriptive statistics for processes used in student spoken responses to the business version of the two assessment tasks ($N = 47$)

Process	ILST ($n = 23$)			MMST ($n = 24$)		
	Mean	Median	Range	Mean	Median	Range
Material	1.26	1.00	0 – 4	1.17	1.00	0 – 3
Mental	1.17	1.00	0 – 6	0.79	0.00	0 – 5
Verbal	0.22	0.00	0 – 2	0.17	0.00	0 – 1
Behavioral	0.48	0.00	0 – 3	0.29	0.00	0 – 1
Relational: Identifying	2.96	3.00	0 – 6	3.58	4.00	1 – 7
Relational: Attributive	1.09	1.00	0 – 3	1.17	1.00	0 – 4
Existential	0.26	0.00	0 – 2	0.71	1.00	0 – 3
Total	7.39	8.00	3 – 10	7.88	8.00	4 – 17

Students in both the ILST and the MMST group used material and attributive relational processes the third and the fourth most frequently. This proportion relative to the total processes resembles that of the input lecture. There were non-significant associations between students' use of mental processes and their assigned task types $\chi^2(6) = 3.42$, $p < .754$, and attributive relational processes and the task types $\chi^2(4) = 1.11$, $p < .892$. In this sense, both task types seemed to succeed in eliciting students' choices of processes appropriate for representing key information required by the task.

First, students in both the ILST and the MMST group used a mental process approximately once, on average, in their spoken responses, $M = 1.17$, $Mdn = 1.00$, and *range* = 0 - 6 for the ILST; $M = 0.79$, $Mdn = 0.00$, and *range* = 0 - 5 for the MMST. For instance, Student 3.24, one of the 12 ILST students who used a mental process once or more, used a mental process, "to choose," to describe the nature of a consumer decision-making process, as shown in the following quote, "… so the consumers must **choose** the products according to their feature …" (Student 3.24, ILST, emphasis added). Similarly, Student 1.9, one of the 11 MMST students who used a mental process once or more, also used "to decide" to describe a thinking process of consumers, as shown in the following quote, "[r]ight now, the

customer decision to buy some product, they **decide** by the high quality of the product …" (Student 1.9, MMST, emphasis added). The linguistic characteristics of the input lecture seemed to influence the choices of language in student spoken responses, because what is explained in the oral responses derives from what is taught in the input lecture.

Second, students in both the ILST and the MMST group also used an attributive relational process approximately once, on average, in their spoken responses, *M* = 1.09, *Mdn* = 1.00, and *range* = 0 - 3 for the ILST; *M* = 1.17, *Mdn* = 1.00, and *range* = 0 - 4 for the MMST. For example, Student 3.7, one of the 16 ILST students who used an attributive relational process once or more, used the process "to be" to describe a "product," as shown in the following quote, "… But if it [a product] is not [working the way it should], then the product **is** unreliable …" (Student 3.7, ILST, emphasis added). Likewise, Student 2.15, one of the 16 MMST students who used an attributive relational process once or more, used the process "to be" to describe the characteristics of features as a factor of product quality, as shown in the following quote, "… It [a feature] **is** more important and easy to use" (Student 2.15, MMST, emphasis added). The types of meaning made in the input lecture such as descriptions appeared to directly affect those in students' spoken response and accordingly, their choice of language representing such meanings.

The moderately involved level of knowledge of mental and attributive relational processes manifested in students' production was consistent with their moderately perceived level of using these process types, based on the item-level mean ratings of the interactional authenticity questionnaire on a six-point Likert-scale, 1) mental processes: *M* = 2.61, *SD* = 1.46 for the ILST group, and *M* = 3.11, *SD* = 1.29 for the MMST group, and 2) attributive relational processes: *M* = 4.33, *SD* = 1.37 for the ILST group, and *M* = 4.47, *SD* = 1.31 for

the MMST group. Different from the mismatch between the perceived involvement of knowledge of material processes and its actual use in student production discussed in sub-section 4.3.2.1., students' perception of using their knowledge of mental and attributive relational processes corresponded to the use of such knowledge manifested in student spoken responses.

4.3.2.3. Summary

All in all, the "what" of the input lecture, represented in processes, affected students' linguistic choices in their spoken responses. The biology input lecture included many occurrences of material processes (48.8%) such as "to sharpen" and "to pull" to describe the physical actions of animals who use tools. Students who took the biology version of the ILST and the MMST also chose material processes frequently, on average, approximately four times out of eight to nine processes in total, when they spoke about the animal examples in their explanation of two definitions of tools. However, this highly frequent use of material processes was not consistent with the only moderate degree of perception based on the item-level meaning ratings of interactional authenticity regarding language knowledge.

On the other hand, mental processes such as "to determine" and "to choose" and attributive relational processes such as "to be" were often used in the business input lecture (19.2% for mental processes; 15.4% for attributive relational processes) to refer to the thinking process involved in consumer decision-making and describe the characteristics of factors to consider in such a process. A similar tendency was found in student spoken responses for both the ILST and the MMST group. Students in both groups used mental and attributive relational processes each, on average, once out of eight processes in total. They

also perceived a moderate use of these two processes according to their ratings of interactional authenticity regarding language knowledge. Student perception and actual use in production matched.

These findings suggest that the topics of input lectures and the types of language used in the lectures make a difference in language choices in student responses. This has an implication for test developers, especially the design of a task-specific rating scale descriptor for content of speech samples. Frost et al. (2012) used an intuitively constructed list of key points of the input texts in validating part of their content-related ratings scale. On the other hand, ideational comparison between input texts and spoken responses will possibly provide a more language-based measure to evaluate the expression of content.

4.3.3. Use of Exophoric Reference for Task Types: The ILST and the MMST

This sub-section addresses the third part of research question three regarding the use of pronouns as exophoric reference (something outside of the text itself) between the two task types, the ILST and the MMST, and also across the three assigned task scores, low, moderate, and high. Forty-six students who took the ILST and 47 who took the MMST provided a one-minute spoken response to their assigned task. These spoken responses were transcribed and pruned, and pronouns as exophoric reference were identified (refer to 3.6.3.3.). The three types are: 1) “we/our/us” referring to the speaker and the imagined audience, 2) “you/your/you” referring to the imagined audience, and 3) “I” referring to the speaker.

Each pronoun used as exophoric reference in each spoken response was counted, and these counts were compared across the groups. The first part of this sub-section presents a

comparison of this linguistic feature used in the ILST and the MMST spoken responses. Then, the same kind of comparison across assigned task scores follows in the second sub-section. The final sub-section provides a summary.

4.3.3.1. Comparisons between the ILST and the MMST

To compare the use of pronouns as exophoric reference among students who took the ILST and the MMST, descriptive statistics for the three types of exophoric pronouns used in student spoken responses were computed, and are presented below in Table 33. As a student spoken response was limited to one minute, many instances of pronouns as exophoric reference were not expected, and this expectation was consistent with the less frequent use of this linguistic feature by students in both the ILST and the MMST group.

Table 33. Descriptive statistics for pronouns as exophoric reference used in student spoken responses to the ILST and the MMST ($N = 93$)

Pronouns as	ILST ($n = 46$)			MMST ($n = 47$)		
Exophoric Reference	Mean	Median	Range	Mean	Median	Range
We/our/us	0.83	0.00	0 – 10	0.72	0.00	0 – 6
You/your/you	0.20	0.00	0 – 2	0.38	0.00	0 – 4
I	0.07	0.00	0 – 1	0.19	0.00	0 – 5
Total	1.09	0.00	0 – 10	1.30	1.00	0 – 7

On average, students in the ILST group used approximately one of the three types of pronouns in their responses, $M = 1.09$, $Mdn = 0.00$, and *range*: 0 – 10, and those in the MMST group used this linguistic feature slightly more than one time, $M = 1.30$, $Mdn = 1.00$, and *range*: 0 – 7. This finding seems to match the ILST student perception of low use of this feature in their spoken responses (see Section 4.2.1.1.) and the MMST student perception of moderate use (see Section 4.2.2.1.).

To statistically compare the use of pronouns as exophoric reference, a contingency table for this linguistic feature was created and a Pearson's chi-square test was conducted. Table 34 below presents the number of students who used pronouns as exophoric reference depending on their assigned assessment task.

Table 34. Contingency table showing the number of students who used pronouns as exophoric reference depending on the assigned assessment task ($N = 93$)

		Task		
		ILST	MMST	Total
Pronouns as Exophoric Reference	Yes	18 (39.1%)	30 (63.8%)	48 (51.6%)
	No	28 (60.9%)	17 (36.2%)	45 (48.4%)
	Total	46 (100.0%)	47 (100.0%)	93 (100.0%)

More than 60% of the MMST students (63.8%) used pronouns as exophoric reference to connect themselves with the imagined audience, although less than 40% of the ILST students (39.1%) used this linguistic feature.

There was a statistically significant association between the type of the assessment tasks and the use of pronouns as exophoric reference $\chi^2(1) = 5.68$, $p < .017$. Based on the odds ratio, the odds of students using pronouns as exophoric reference were 2.75 times higher if they took the MMST than the ILST. Additional visual elements in the MMST such as a professor moving in the input video seemed to simulate a classroom setting more successfully, and provide a sense of the imagined audience over the other side of the computer screen to which students spoke during the task. This may have encouraged students to choose pronouns such as "we" and "you" to address those who would potentially listen to their responses.

4.3.3.2. Comparisons between the ILST and the MMST per task score

It is expected that students who achieve a higher score on the assigned assessment task will use the targeted language knowledge, defined in the construct, more frequently during their task performance. To statistically compare the use of pronouns as exophoric reference such as "we" and "you" between the two assessment tasks and also across the assigned task scores (low, moderate, and high), a contingency table for this linguistic feature was created, and a three-way loglinear analysis was conducted. Table 35 below presents the number of students who used pronouns as exophoric reference depending on their assigned assessment task and task score.

Table 35. Contingency table showing the number of students who used pronouns as exophoric reference depending on the assigned assessment task and task score ($N = 93$)

		Task						Total
		ILST			MMST			
		Low	Moderate	High	Low	Moderate	High	
Pronouns as Exophoric Reference	Yes	3	10	5	9	16	5	48
		33%	45%	33%	82%	59%	56%	52%
	No	6	12	10	2	11	4	45
		67%	55%	67%	18%	41%	44%	48%
	Total	9	22	15	11	27	9	93
		100%	100%	100%	100%	100%	100%	100%

To test for significance of the associations, a three-way loglinear analysis, the three-category version of a chi-square test, was used. This analysis produced a final model that retained only the one-way effects. The likelihood ratio of this model was χ^2 (0) = 0, p = 1. This indicated that all the two-way interactions and the three-way interaction collectively were not significant, χ^2 (7) = 10.88, p < .144, and specifically, the three-way interaction alone was also not significant, χ^2 (2) = 1.99, p < .369. This analysis revealed that the association between the type of assessment tasks and the use of pronouns as exophoric reference in student spoken

responses did not differ significantly with the assigned task score. Contrary to expectations, the degree of students' use of the targeted language knowledge did not significantly increase with the higher assigned task scores. It seems that an ability of being aware of audience and choosing language accordingly does not vary across different English proficiency levels.

4.3.3.3. Summary

The focal linguistic feature in this section was exophoric pronouns, referring to something outside of the text itself, and categorized into three types: 1) pronouns referring to the speaker and the imagined audience such as "we," "our," and "us," 2) pronouns referring to the imagined audience such as "you," and "your," and 3) pronouns referring to the speaker such as "I." The use of these pronouns as exophoric reference was compared between the two assessment tasks, the ILST and the MMST.

Students in the ILST group used approximately one of the three types of pronouns in their responses, $M = 1.09$, $Mdn = 0.00$, and *range*: 0 – 10, and those in the MMST group used this linguistic feature slightly more than one time, $M = 1.30$, $Mdn = 1.00$, and *range*: 0 – 7. This infrequent occurrence was expected as a student spoken response was limited to a minute. This low usage was also consistent with a low level of student perception of using this language feature, according to the item-level mean ratings indicated in the interactional authenticity regarding language knowledge.

Of the students who took the MMST, 63.8%, used one of the three types of pronouns as exophoric reference once or more, while only 39.1% of the students who took the ILST did so. This group difference was statistically significant $\chi^2(1) = 5.68$, $p < .017$. Based on the odds ratio, the odds of students using pronouns as exophoric reference were 2.75 times

higher if they took the MMST than the ILST. However, there was no statistically significant group difference across the assigned task scores. It appears that visual information in the MMST helped students to have a better sense of the audience to their responses and use pronouns such as "we" and "you" to address those imaginary listeners.

4.3.4. Section Summary

To address the third research question, the transcripts of 46 one-minute spoken responses to the ILST and 47 responses to the MMST were analyzed in terms of a few linguistic features. First, the use of processes, in other words, a grammatical unit representing meaning such as some activity or way of being, was compared between the ILST and the MMST group and across three assigned task scores, low, moderate, and high (research question three, sub-section three). The two target processes, defined in the task construct, were: 1) identifying relational processes (e.g., to be), and 2) existential processes (e.g., there are). Students' use of identifying relational processes was not significantly different between the groups, χ^2 (1) = .79, p < .373, and also across the task scores, χ^2 (2) = 0.73, p < .695. On the other hand, students who took the MMST used existential processes more than those in the ILST group, and this group difference significantly differed depending on the assigned task χ^2 (1) = 4.86, p < .027. It appears that the content visuals included in the MMST may have helped students to understand the classifying nature of concepts being explained in the lecture and to use an appropriate type of language to represent such meaning.

Second, the use of processes between the two versions of the ILST and the MMST, biology and business, was compared (the second sub-research question three). The biology input lecture included many occurrences of material processes such as "to use" and "to

sharpen," and students in both the ILST and the MMST group also used this process type frequently, on average, approximately four times out of eight to nine processes in total. The business input lecture included relatively many occurrences of mental processes such as "to determine" and "to choose" and attributive relational processes such as "to be." A similar tendency was also found from student spoken responses for both the ILST and the MMST group. The content of the input lecture, represented in different types of processes, seems to influence students' linguistic choices in their responses.

Third, the use of pronouns as exophoric reference (something outside of the text itself) such as "we" and "you" was compared between the ILST and the MMST group and across three assigned task scores, low, moderate, and high (the third sub-research question three). Students' use of pronouns as exophoric reference was significantly different between the task types χ^2 (1) = 5.68, $p < .017$. However, a group difference per assigned task score was not significantly different χ^2 (2) = 1.99, $p < .369$.

In conclusion, among the target language knowledge which was expected during task completion, both the ILST and the MMST equally functioned well in eliciting student knowledge of identifying relational processes and processes appropriate to the lecture content. On the other hand, knowledge of existential processes and pronouns as exophoric reference was tapped more successfully by the MMST than the ILST. Additional channels of the input lecture such as content visuals and a moving professor in the video seem to facilitate an elicitation of the construct-relevant language knowledge in student performance.

4.4. Strategic Competence Involved during Task Performance

This section reports on the results of the fourth research question: How is strategic competence involved in completing the two assessment tasks—the integrated listening-speaking task (ILST) and the multimedia-mediated speaking task (MMST)? How different is this elicited strategic competence between the tasks and across assigned task scores? Strategic competence in the current study is operationalized as *reported* actions and thought processes of the participants. The data consist of transcripts of the 20 strategic behavior reports from stimulated recall of students who took either the ILST or the MMST. Counts of strategy types identified in each of the strategic behavior reports were used to interpret the types and extent of elicited strategic competence of the examinees in the two assessment tasks.

In the first sub-section, I will present the descriptive statistics showing the extent to which students reported the involvement of their strategic competence in accomplishing the integrated listening-speaking task (ILST) and the multimedia-mediated speaking task (MMST), and compare the counts of strategy types between the two groups. To assist in the interpretations, quotes from the transcripts of strategic behavior reports are also provided. In the second sub-section, I will present the same kind of descriptive statistics across the three assigned task scores (low, moderate, and high) and compare them between the two assessment tasks. The final sub-section provides a summary.

4.4.1. Comparisons Between the ILST and the MMST

Ten students who took the integrated listening-speaking task (ILST) and another ten who took the multimedia-mediated speaking task in the Spring 2013 administration

participated in a stimulated recall interview, and reported what they had been thinking during task performance. These strategic behavior reports were transcribed and coded using a modified version of Swain et al.'s (2009) strategy scheme (refer to Appendix I). The number of each strategy identified in each strategic behavior report was counted and compared across the groups. Table 36 below presents the descriptive statistics for the five main categories of strategy reported in students' strategic behavior reports.

Table 36. Descriptive statistics for categories of strategy reported in strategic behavior reports ($N = 20$)

Strategy Category	ILST ($n = 10$)			MMST ($n = 10$)		
	Mean	Median	Range	Mean	Median	Range
Approach	1.30	1.50	0 – 2	1.30	1.00	0 – 4
Communication	3.20	3.00	1 – 6	2.80	3.00	0 – 6
Cognitive	4.90	5.00	2 – 8	7.90	8.00	2 – 13
Metacognitive	5.10	5.00	2 – 8	4.10	4.00	1 – 7
Affective	0.90	1.00	0 – 3	0.00	0.00	0
Total	15.40	15.00	11 – 24	16.10	14.00	10 – 28

On average, 15 to 16 instances of strategy use were reported by students in both groups (M = 15.40 for the ILST and M = 16.10 for the MMST). Among the five strategy categories, students in both groups used many instances of communication, cognitive and metacognitive strategies. These three categories of strategies together constitute almost all of the occurrences of reported strategies. This tendency is comparable to the strategic behavior reports from Swain et al. (2009). In their study in which they used an integrated listening-speaking task similar to the ILST, they also found a high number of communication, cognitive, and metacognitive strategy use.

To examine detailed strategy use in each of the three most often reported strategy categories, the descriptive statistics for sub-categories of strategies under each were computed and are presented below (Tables 37 – 39).

Communication strategies are defined as conscious plans for solving a linguistic problem in order to reach a communicative goal and, in the current study, include nine sub-categories: 1) paraphrasing, 2) approximating, 3) linking to prior experiences/knowledge, 4) reviewing notes, 5) referring to notes, 6) organizing thoughts, 7) guessing, 8) rehearsing, and 9) elaborating to fill time. Table 37 presents the descriptive statistics for the sub-categories of communication strategies reported in strategic behavior reports. Among the nine sub-categories, more than half of the students in both the ILST and the MMST group did not report any instances of the sub-categories 1, 2, 7, 8 and 9 (*Mdn* = 0.00 for both groups), and these are excluded from the table for less dense visual presentation.

Table 37. Descriptive statistics for the sub-categories of communication strategies reported in strategic behavior reports (*N* = 20)

Communication Strategies		ILST (*n* = 10)			MMST (*n* = 10)		
		Mean	Median	Range	Mean	Median	Range
3	Linking to prior experience/knowledge	0.80	1.00	0 – 3	0.40	0.00	0 – 1
4	Reviewing notes	0.90	1.00	0 – 2	0.70	1.00	0 – 2
5	Referring to notes	0.40	0.00	0 – 1	1.00	1.00	0 – 2
6	Organizing thoughts	0.60	1.00	0 – 1	0.30	0.00	0 – 1
	Total	3.20	3.00	1 – 6	2.80	3.00	0 – 6

Overall, students in both groups reported approximately three instances of using communication strategies during their task performance (*M* = 3.20, *Mdn* = 3.00, and *range* = 1 - 6 for the ILST; *M* = 2.80, *Mdn* = 3.00, and *range* = 0 - 6 for the MMST). Especially, the sub-categories of reviewing and referring to notes were relatively frequently used. This finding from strategic behavior reports is consistent with a highly perceived level of using the same sub-categories based on student responses to the interactional authenticity questionnaire (refer to sub-sections 4.2.1.2. and 4.2.2.2.).

Cognitive strategies refer to the manipulation of the target language to understand and produce language, and, in this study, consist of seven sub-categories: 1) anticipating the content, 2) anticipating the structure, 3) anticipating the question, 4) using imagery, 5) using notes to organize or remember information, 6) translating, and 7) inferencing. Among these seven sub-categories, more than half of the students in both the ILST and the MMST group did not report any instances of the sub-categories 2, 3, 6, and 7 (*Mdn* = 0.00 for both groups), which, therefore, are not included in Table 38 for visually better presentation.

Table 38. Descriptive statistics for the sub-categories of cognitive strategies reported in strategic behavior reports (*N* = 20)

Cognitive Strategies	ILST (*n* = 10)			MMST (*n* = 10)		
	Mean	Median	Range	Mean	Median	Range
1 Anticipating the content	1.80	1.50	1 – 4	3.50	3.00	0 – 6
4 Using imagery	0.50	0.00	0 – 2	1.40	1.00	0 – 4
5 Using notes	1.70	2.00	0 – 4	2.10	1.50	0 – 6
Total	4.90	5.00	2 – 8	7.90	8.00	2 – 13

Students in the MMST group reported more instances of using cognitive strategies overall (*M* = 7.90, *Mdn* = 8.00, and *range* = 2 - 13) than did those in the ILST group (*M* = 4.90, *Mdn* = 5.00, and *range* = 2 - 8). The sub-categories 1, 4, and 5 were also reported more frequently by students in the MMST group. Specifically, the strategy of anticipating the content (sub-category 1) was reported more often by the MMST students (*M* = 3.50, *Mdn* = 3.00, and *range* = 0 - 6) than by the ILST students (*M* = 1.80, *Mdn* = 1.50, and *range* = 1 - 4). The instances of using the strategy of anticipating the content by the MMST students were usually triggered by content visuals included in the video input lecture. For instance, Student 2.1, one of the students who took the MMST, predicted the topic of the lecture when exposed to a picture illustrating the introduction of the lecture:

I saw the two images on the screen for the red car and the blue cars [refer to Figure 6]. I was thinking that this lecture might be related with the car things. (Student 2.1) (supplemental information added in brackets)

Figure 22. The first content visual used in the MMST

These frequent reports of anticipating the content strategy do not match only a moderately perceived use of this strategy (refer to sub-sections 4.2.1.2. and 4.2.2.2.).

In addition, students in the MMST group reported relatively more instances of the strategy of using imagery (M = 1.40, Mdn = 1.00, and *range* = 0 - 4) than did those in the ILST group (M = 0.50, Mdn = 0.00, and *range* = 0 - 2). This finding is consistent with a statistically significant group difference in the student perception of using this strategy type ($F(1, 71) = 4.86$, $p = .031$, discussed in sub-section 4.2.3.2.). Content visuals in the MMST input lecture seemed to facilitate the use of imagery in understanding and remembering information. For example, when the business lecture explained how the role of reliability and features had changed in consumer decision making, a picture of a seesaw was presented to illustrate the relative importance of the two factors (Figure 23).

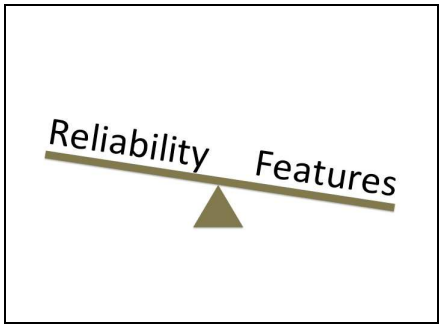


Figure 23. The last content visual used in the MMST

The following example shows how this illustration was used by Student 4.1, one of the MMST students, to comprehend the lecture:

> So in this case, the balance [seesaw], the balance, the, the, the illustration of balance help me, like to help me to confirm what, what I, what I'm listening like it's not deciding factor, factor, like reliability is not deciding factor, factors, and yeah, the, the visual aids help me to, to confirm this, and like pay more attention to the second one [features] like, the second one is more important than, than the previous one [reliability]. So, yeah, so the visual aids help me more to, to confirm what I'm listening. (Participant 4.1) (supplemental information added in brackets)

The MMST participants implemented a strategy of using the content visuals to pay attention to and comprehend key information of the input lecture.

The strategy of using notes was frequently reported by students in both groups (M = 1.70, Mdn = 2.00, and *range* = 0 – 4 for the ILST; M = 2.10, Mdn = 1.50, and *range* = 0 - 6). This report is consistent with a high student perception of using this strategy type in both groups (refer to sub-sections 4.2.1.2. and 4.2.2.2.). Highly perceived degrees of using all three strategies related to note taking, reviewing, and referring to notes under the category of communication strategies and using notes under the category of cognitive strategies matched frequent occurrences of using these strategies in strategic behavior reports.

Metacognitive strategies in the current study involve eight sub-categories: 1) setting goals, 2) identifying the purpose of the task, 3) planning, 4) monitoring, 5) self-correcting, 6) evaluating the content of what was heard/seen, 7) evaluating performance while speaking, and 8) evaluating language production after completing a task. Table 39 below presents the descriptive statistics for the sub-categories of metacognitive strategies reported by students in both the ILST and the MMST group. Among the eight sub-categories, more than half of the students in both groups did not report any occurrences of the sub-categories 4, 5, and 7 (*Mdn* = 0.00 for both groups), and these are not included in the table for clearer organization.

Table 39. Descriptive statistics for the sub-categories of metacognitive strategies reported in strategic behavior reports (*N* = 20)

Metacognitive Strategies		ILST (*n* = 10)			MMST (*n* = 10)		
		Mean	Median	Range	Mean	Median	Range
1	Setting goals	0.30	0.00	0 – 1	0.40	0.00	0 – 1
2	Identifying the task purpose	0.70	0.50	0 – 2	0.60	0.00	0 – 2
3	Planning	1.60	1.50	1 – 3	1.30	1.00	0 – 2
6	Evaluating the content	0.80	1.00	0 – 2	0.30	0.00	0 – 2
8	Evaluating language production	0.60	1.00	0 – 1	0.60	0.50	0 – 2
	Total	5.10	5.00	2 – 8	4.10	4.00	1 – 7

Students in both the ILST and the MMST group reported approximately five instances of using metacognitive strategies during their task performance (*M* = 5.10, *Mdn* = 5.00, and *range* = 2 - 8 for the ILST; *M* = 4.10, *Mdn* = 4.00, and *range* = 1 - 7 for the MMST). The sub-category of planning was relatively frequently used by students in both groups (*M* = 1.60, *Mdn* = 1.50, and *range* = 1 - 3 for the ILST; *M* = 1.30, *Mdn* = 1.00, and *range* = 0 - 2 for the MMST), although this strategy was perceived as only moderately used in the interactional authenticity questionnaire (refer to sub-sections 4.2.1.2. and 4.2.2.2.). On the other hand, students both reported and perceived using the sub-category of evaluating

language production (*M* = 0.60, *Mdn* = 1.00, and *range* = 0 - 1 for the ILST; *M* = 0.60, *Mdn* = 0.50, and *range* = 0 - 2 for the MMST) to a relatively moderate degree. Interestingly, students perceived a high to moderate use of the strategy of setting goals in their interactional authenticity questionnaire ratings; on the contrary, this type of strategy was not reported frequently in strategic behavior reports (*M* = 0.30, *Mdn* = 0.00, and *range* = 0 - 1 for the ILST; *M* = 0.40, *Mdn* = 0.00, and *range* = 0 - 1 for the MMST). These mismatches between student perception and reports of actual use in performance may have occurred, as metacognition is high-level thinking and may not be consciously noticed and remembered.

4.4.2. Comparisons Between the ILST and the MMST per Task Score

To examine how student strategic competence was involved across three assigned task scores (low, moderate, or high), mean counts of the five categories of strategy reported by students in the ILST and the MMST group were calculated per task score, and are presented in the table below.

Table 40. Mean reported counts of the five strategy categories per assigned task score (*N* = 20)

Strategy Category	ILST (*n* = 10)			MMST (*n* = 10)		
	Low (*n* = 3)	Moderate (*n* = 4)	High (*n* = 3)	Low (*n* = 3)	Moderate (*n* = 6)	High (*n* = 1)
Approach	1.00	1.25	1.67	0.33	1.67	2.00
Communication	3.33	3.00	3.33	2.67	2.67	4.00
Cognitive	4.33	4.00	6.67	8.00	7.83	8.00
Metacognitive	3.33	5.50	6.33	4.00	4.00	5.00
Affective	1.00	0.25	1.67	0.00	0.00	0.00
Total	13.00	14.00	19.67	15.00	16.17	19.00

Students, on average, reported more instances of strategies in total as the assigned task scores increased, and similar increasing patterns were found in most of the strategy

categories for both groups. The exceptions were the mean reported counts of the communication strategies for the ILST group and those of the cognitive strategies for the MMST group that were similar across the three task scores. Noticeably, the mean occurrences of the cognitive strategies reported by the MMST students whose task scores were low and moderate (M = 8.00 for low; M = 7.83 for moderate) differed largely from those by the ILST students assigned to the same task score groups (M = 4.33 for low; M = 4.00 for moderate) (refer to Figure 24).

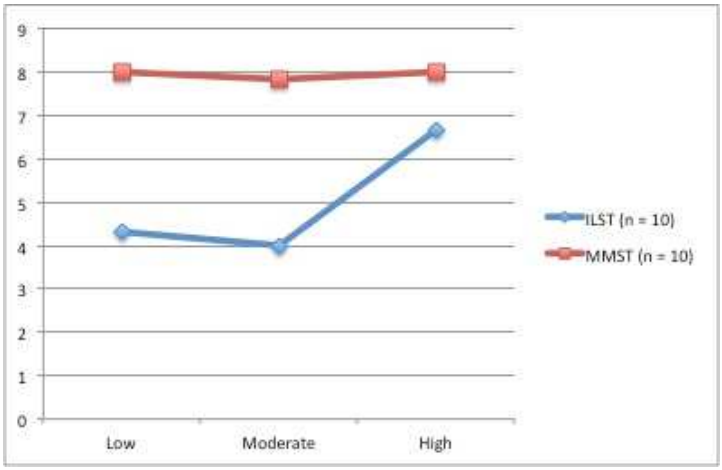


Figure 24. The ILST and MMST mean reported counts of the cognitive strategies among the three score groups (N = 20)

To examine if this notable score group difference is also identified for the sub-categories of the cognitive strategies, reported counts of each sub-strategy were computed per task score group, and are presented in Table 41. Each sub-strategy showed different patterns of the mean reported counts across the three assigned task score groups.

Table 41. Mean reported counts of the sub-categories of cognitive strategies per assigned task score ($N = 20$)

	Cognitive Strategies	ILST ($n = 10$)			MMST ($n = 10$)		
		Low ($n = 3$)	Moderate ($n = 4$)	High ($n = 3$)	Low ($n = 3$)	Moderate ($n = 6$)	High ($n = 1$)
1	Anticipating the content	2.00	2.00	1.33	3.67	3.50	3.00
4	Using imagery	0.00	0.50	1.00	1.33	1.00	4.00
5	Using notes	1.00	1.25	3.00	1.33	2.67	1.00
	Total	4.33	4.00	6.67	8.00	7.83	8.00

The mean reported counts of the first sub-strategy, anticipating the content, for both the ILST and the MMST group across the assigned scores slightly decreased to a similar degree, but overall, the mean reported counts for the MMST group were higher than those for the ILST group (refer to Figure 25). It appears that content visuals in the MMST input lectures promoted more frequent anticipation of the lecture content regardless of students' English proficiency.

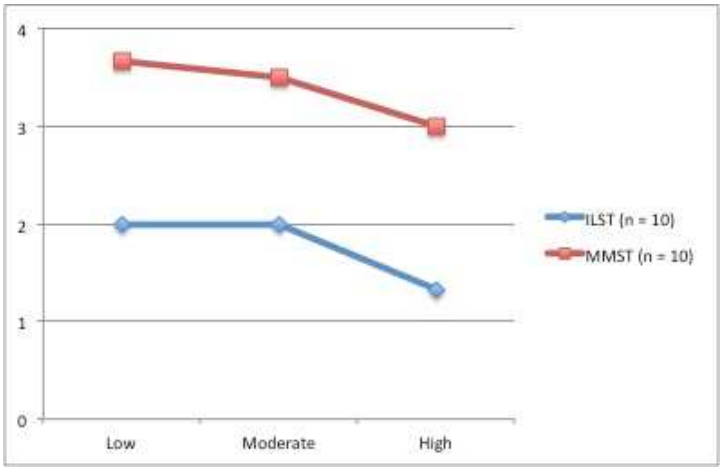


Figure 25. The ILST and MMST mean reported counts of sub-strategy of anticipating the content among the three score groups ($N = 20$)

The sub-strategy of using imagery for both the ILST and the MMST group showed a generally increasing pattern of mean reported counts across the assigned task scores (see

Figure 26). However, the mean reported counts for the MMST group increased more sharply than those for the ILST group.

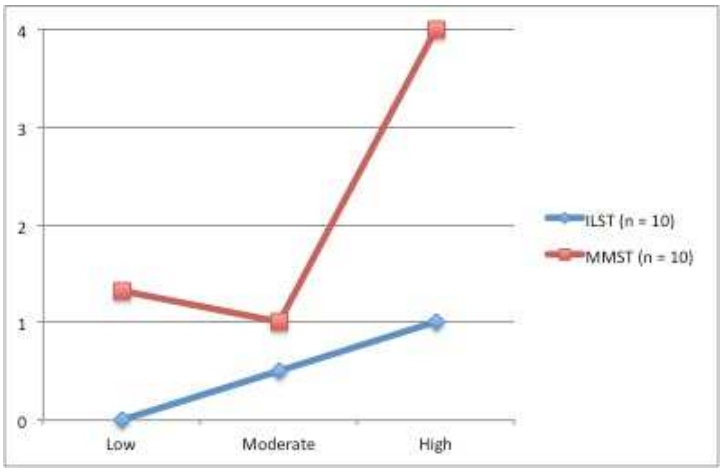


Figure 26. The ILST and MMST mean reported counts of sub-strategy of using imagery among the three score groups ($N = 20$)

More proficient students may use imagery more frequently to understand and remember information, or this finding may simply be due to the small sample size (the MMST-High group has only one student as indicated in Table 41).

The sub-strategy of using notes showed a different pattern between the ILST and the MMST group. Whereas the mean reported counts for the ILST group increased across the three task scores, those for the MMST group slightly decreased overall with a high peak for the MMST-Moderate group (refer to Figure 27).

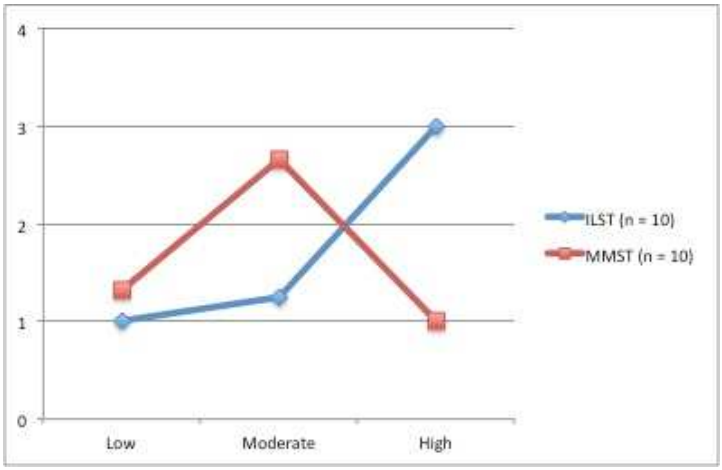


Figure 27. The ILST and MMST mean reported counts of sub-strategy of using notes among the three score groups ($N = 20$)

This unique pattern may truly be that of the population, or just due to a sampling error. An analysis with more reported strategy data from a larger number of samples will confirm or deny the pattern of mean counts of this sub-strategy.

4.4.3. Section Summary

To address the fourth research question, the transcripts of stimulated recall interviews from ten students who took the ILST and another ten who took the MMST in the Spring 2013 administration were coded using a modified version of Swain et al.'s (2009) strategy categories, and the types of strategies identified in the transcripts were analyzed. Among the five main strategy categories, students in both the ILST and the MMST group reported a frequent use of communication, cognitive, and metacognitive strategies. Under the category of communication strategies, students in both groups reported a frequent use of the following four sub-categories: 1) linking to prior experience/knowledge, 2) reviewing notes, 3) referring to notes, and 4) organizing thoughts.

Under the category of cognitive strategies, the following three sub-categories were frequently reported as being used: 1) anticipating the content, 2) using imagery, and 3) using notes. Specifically, the first two sub-categories of strategy were more frequently reported by the MMST students (*M* = 3.50, *Mdn* = 3.00, and *range* = 0 - 6 for the strategy of anticipating the content; *M* = 1.40, *Mdn* = 1.00, and *range* = 0- 4 for the strategy of using imagery) than the ILST students (*M* = 1.80, *Mdn* = 1.50, and *range* = 1 - 4 for the fist sub-category; *M* = 0.50, *Mdn* = 0.00, and *range* = 0- 2 for the second sub-category). Content visuals included in the MMST input lectures seemed to facilitate student use of these two sub-categories.

Under the category of metacognitive strategies, students in both the ILST and the MMST groups reported a relatively frequent use of the following five sub-categories of strategy: 1) setting goals, 2) identifying the task purpose, 3) planning, 4) evaluating the content, and 5) evaluating language production. Among these five, the strategy of setting goals was reported least by students in both groups (*M* = 0.30, *Mdn* = 0.00, and *range* = 0 -1 for the ILST; *M* = 0.40, *Mdn* = 0.00, and *range* = 0 -1 for the MMST). This low reported frequency is possibly due to high-level thinking involved in implementing this strategy that may not be consciously noticed and remembered.

The mean reported counts of the five main strategy categories generally increased across the assigned task scores (low, moderate and high). Interestingly, the mean occurrences of the cognitive strategies reported by the MMST-low and MMST-moderate students (*M* = 8.00 for low; *M* = 7.83 for moderate) were considerably different from those by the ILST students assigned to the same task score groups (*M* = 4.33 for low; *M* = 4.00 for moderate). However, the patterns of the mean reported counts across the assigned task scores varied for each of the sub-categories of cognitive strategies. First, the mean reported counts for the

strategy of anticipating the content slightly decreased across three task scores, although the mean counts reported by students in the MMST group were larger throughout than those by students in the ILST group. On the other hand, the mean reported counts for the strategy of using imagery for both groups increased across the assigned task scores, but the increase of those for the MMST group was sharper. Lastly, the mean reported counts for the strategy of using notes across the assigned scores showed an interesting pattern, as those for the ILST group increased sharply but those for the MMST group showed a relatively downward pattern with a high peak for the moderate score group. More data from a larger number of samples seem to be required to confirm or deny this pattern.

All in all, both the ILST and the MMST equally worked well in facilitating student use of the communication and metacognitive strategies defined in the task construct with a relatively increasing magnitude across the assigned task scores. On the other hand, it seems that the MMST was better at promoting the use of cognitive strategies than the ILST was with the same great magnitude across the task scores. Content visuals appear to be frequently utilized and assist students in understanding and remembering key information of the input lecture regardless of student English proficiency.

4.5. Chapter Summary

This chapter presented the findings of the data analyses for answering each of the four research questions and discussed the interpretations of the findings. The first research question concerned the degree of situational authenticity of the ILST and the MMST. Students perceived a moderate degree of situational authenticity of the ILST overall. The mean ratings on individual questionnaire items ranged from a moderate to a high level of

situational authenticity of the ILST. This student perception empirically supports the widely accepted notion of authenticity in integrated assessment tasks (Plakans, 2013; Weir, 1990). Students also perceived the overall degree of situational authenticity of the MMST as similar to that of the ILST. The mean ratings on individual questionnaire items for the situational authenticity of the MMST ranged from moderate to high, and did not differ from those for the ILST. The adding of video input to the MMST did not seem to contribute to the enhancement of situational authenticity, as measured by the questionnaire eliciting students' perceptions, or a ceiling effect in student ratings may have resulted in non-significant differences between the groups.

The second research question concerned the degree of interactional authenticity of the ILST and the MMST. Students who took either the ILST or the MMST perceived a similar degree of interactional authenticity regarding language knowledge for both assessment tasks. Similar degrees of perception for the two tasks were also found across the assigned task scores. For strategic competence, students in both groups perceived a similar degree of interactional authenticity, except for the strategy of using imagery that students in the MMST group perceived as higher than those in the ILST group. It appears that content visuals included in the MMST encouraged students to use imagery more frequently for understanding and remembering information given by the input lecture. However, perceived degrees of interactional authenticity regarding strategic competence across the assigned task scores were similar for the two assessment tasks. Generally, students indicated that they perceived both the ILST and the MMST performed similarly in tapping students' language ability consisting of language knowledge and strategic competence.

The third research question concerned elicited language knowledge manifested in student spoken responses to the ILST and the MMST. Both the ILST and the MMST functioned well in eliciting the use of the following features of the target language knowledge: identifying relational processes, material processes for the biology version, and mental and attributive relational processes for the business version. However, existential processes and pronouns as exophoric reference were elicited more successfully by the MMST than by the ILST. According to these linguistic analyses of student spoken responses, the extent of involvement of examinees' language knowledge required to accomplish an assessment task, in other words, interactional authenticity regarding language knowledge, was higher for the MMST. However, the lack of association between these features and raters' judgments about overall level of performance indicate that these features may not have been judged as important for distinguishing among examinees' performance.

The fourth research question concerned strategic competence involved in accomplishing the ILST and the MMST. Students in both the ILST and the MMST group reported a frequent use of communication, cognitive, and metacognitive strategies. The mean reported counts of communication and metacognitive strategies were relatively similar between the two groups. However, students in the MMST group reported a more frequent use of cognitive strategies than those in the ILST group. Specifically, the MMST students reported more instances of the strategies of anticipating the content and using imagery than the ILST students. Often, these two frequent strategy uses were facilitated by content visuals included in the MMST. Additional visual channels appeared to encourage students to use the strategies of anticipating the content and using imagery for good comprehension and recall. Across the assigned task scores (low, moderate and high), the mean reported counts generally

showed an upward pattern. However, students in the MMST group obtaining low and moderate task scores, on average, reported high mean counts of the cognitive strategies similar to those of the MMST students obtaining high task scores. It seems that students in the MMST group used this strategy type regardless of their English proficiency and content visuals in the input lectures may have facilitated a frequent use for all students.

CHAPTER 5. CONCLUSION

This chapter begins with a summary of the main findings from this dissertation research. In the summary, the answers to the four main research questions are provided. Next, theoretical, practical, and methodological implications are presented. The sub-section for theoretical implications discusses the authenticity of integrated listening-speaking tasks. The sub-section for practical implications describes the function of visuals in integrated tasks and the design of language-based rating scale descriptors for assessing the content of responses to integrated tasks. The sub-section for methodological implications examines an extensive analysis of elicited language ability from the interactionalist perspective and a new discourse analytic approach for studying language elicited from second language assessments.

The third and fourth sections discuss limitations of the study and suggestions for future research, respectively. Limitations of the current study pertain to interpretation of student perception data from questionnaires, equivalency of group characteristics, and the sample size in each analytic unit. Suggestions for future research concern 1) the modification of the situational authenticity questionnaire, 2) authenticity studies on other types of integrated tasks, 3) inclusion of stakeholder groups other than students in the analysis of situational authenticity, 4) consideration of various other lecture topics in examining the relationship between types of language used in the input lectures and language choices made by examinees in their responses, and 5) the function of nonverbal communication features in the video input lectures to student task performance. Lastly, the chapter ends with a conclusion.

5.1. Summary of the Main Findings

In this study, I addressed four research questions. The first two research questions examined how students perceived the situational and interactional authenticity, respectively, of the integrated listening-speaking task (ILST) and the multimedia-mediated speaking task (MMST). The third and fourth research questions investigated the interactional authenticity of the two assessment tasks on the basis of two types of student production data: 1) spoken responses to the tasks for the third research question, and 2) oral reports of strategy use during task performance for the fourth research question.

The first research question concerned the degree of situational authenticity of the ILST and the MMST. Situational authenticity is defined as the perceived relevance of the task characteristics to the features of a specific target language use (TLU) situation (Bachman, 1991). The TLU situation, which the two assessment tasks were designed to simulate, was an academic situation in which an academic concept is explained. The task characteristics can be described following the Bachman and Palmer (2010) framework that covers the setting, rubrics, input, response, and relationship between input and response. The situational authenticity questionnaire (refer to Appendix G), structured around Bachman and Palmer's task characteristics framework and adapted from Liu (2006), was used in this study to investigate to what degree students perceive the characteristics of the ILST and the MMST as relevant to those of the corresponding TLU situation, namely, an explaining/informing task in an academic context.

Forty-six students who took the ILST and 47 who took the MMST responded to the ten items of the situational authenticity questionnaire on a six-point Likert scale. Students perceived a moderate degree of situational authenticity of the ILST overall (M = 4.74, SD =

0.67). The mean ratings on individual questionnaire items ranged from a moderate (e.g., M = 4.39, SD = 1.53 for item 8, problem identification) to a high (e.g., M = 4.96, SD = 1.15 for item 7, sociolinguistic characteristics) level of situational authenticity of the ILST. This student perception empirically supports the widely accepted notion of authenticity in integrated assessment tasks (Plakans, 2013; Weir, 1990).

Students also perceived the overall degree of situational authenticity of the MMST as similar to that of the ILST (M = 4.58, SD = 0.76). These similar perceptions of situational authenticity between the ILST and the MMST were statistically confirmed by the one-way multivariate analysis of variance (MANOVA), Λ = 0.88, $F(10, 62) = 0.82$, p = .614. The mean ratings on individual questionnaire items for the situational authenticity of the MMST ranged from moderate (e.g., M = 4.05, SD = 1.22 for item 9, relationship between input and response) to high (e.g., M = 5.02, SD = 1.05 for item 6, organizational characteristics). Separate univariate ANOVAs for the ten items also showed statistically non-significant differences in perceived situational authenticity between the ILST and the MMST. The MMST, a new type of integrated assessment task specifically designed for the current study, was perceived as having an equal degree of authenticity as the ILST. The adding of video input to the MMST did not seem to contribute to the enhancement of situational authenticity, as measured by the questionnaire eliciting students' perceptions, or a ceiling effect in student ratings may have resulted in non-significant differences between the groups.

The second research question concerned the degree of interactional authenticity of the ILST and the MMST. The interactional authenticity refers to the type and extent of involvement of examinees' language ability in accomplishing an assessment task. The language ability that the ILST and the MMST were designed to elicit was defined using an

interactionalist perspective, and consists of two components: 1) language knowledge, and 2) strategic competence that governs the use of language knowledge in response to a given language use situation. The three critical aspects of target language knowledge of the ILST and the MMST were: 1) identifying relational processes (e.g., "to be" used for giving the entity in question a definite identity), 2) existential processes (e.g., "there are" used for stating something exists), and 3) pronouns as exophoric reference (e.g., "we/our/us" referring to the speaker and the imagined audience). The target strategic competence of the ILST and the MMST included eight target strategies classified into three categories derived from Swain et al. (2009): 1) communication strategies (conscious plans for solving a linguistic problem to reach a communicative goal), 2) cognitive strategies (manipulation of the target language to understand and produce language), and 3) metacognitive strategies (management of organizing, planning, and evaluating).

The interactional authenticity questionnaire (Appendix H) asked students to rate how much the target language knowledge and strategic competence were involved in completing their assigned task, either the ILST or the MMST, on a six-point Likert scale (1 = not at all; 6 = a lot). For the first type of target language knowledge, identifying relational processes, students who took both the ILST and the MMST perceived a relatively moderate use of this knowledge ($M = 4.78$, $SD = 1.31$ for the ILST; $M = 4.30$, $SD = 1.18$ for the MMST). This group difference was not also statistically significant $t(71) = 1.65$, $p = .104$. For the second type of target language knowledge, existential processes, students in both groups perceived a relatively moderate use of this knowledge ($M = 3.83$, $SD = 1.75$ for the ILST; $M = 3.78$, $SD = 1.65$ for the MMST), and this was not statistically different between the groups $t(71) = 0.12$, $p = .901$. For the third type of target language knowledge, pronouns as exophoric reference,

students who took both the ILST and the MMST perceived a relatively less frequent use of this knowledge ($M = 2.58$, $SD = 1.81$ for the ILST; $M = 2.95$, $SD = 1.53$ for the MMST). This group difference was not also statistically significant $t(71) = -0.93$, $p = .358$. Perceived degrees of the three types of target language knowledge involved in both the ILST and the MMST across the assigned task scores (low, moderate, and high) generally showed a relatively upward pattern, and did not significantly differ between the two task groups, $F(2, 67) = 1.22$, $p = .301$ for knowledge of identifying relational processes, $F(2, 67) = 0.76$, $p = .470$ for knowledge of existential processes, and $F(2, 67) = 2.80$, $p = .068$ for knowledge of pronouns as exophoric reference.

For strategic competence, students perceived a relatively moderate degree of interactional authenticity of the ILST overall ($M = 4.44$, $SD = 0.90$). The mean ratings on individual questionnaire items ranged from a moderate (e.g. $M = 3.33$, $SD = 1.90$ for item 9, using imagery) to a high (e.g. $M = 5.14$, $SD = 1.29$ for item 10, using notes) level of interactional authenticity of the ILST. Perceived degrees of the overall and individual aspects of strategic competence involved in the ILST across the assigned task scores (low, moderate, and high) generally showed a relatively upward pattern except the strategy of using imagery that showed similar degrees across the assigned score groups.

Students also perceived the overall degree of interactional authenticity of the MMST regarding strategic competence as similar to that of the ILST ($M = 4.45$, $SD = 0.94$). These similar perceptions between the ILST and the MMST were statistically confirmed by the one-way multivariate analysis of variance (MANOVA), $\Lambda = 0.84$, $F(8, 64) = 1.32$, $p = .174$. The mean ratings on eight individual questionnaire items for the interactional authenticity of the MMST regarding strategic competence ranged from a moderate (e.g., $M = 3.68$, $SD =$

1.51 for item 14, evaluating language production) to a high (e.g., $M = 5.03$, $SD = 1.14$ for item 10, using notes) degree of involvement. Separate univariate ANOVAs for the individual items also showed statistically non-significant differences in perceived interactional authenticity regarding strategic competence between the ILST and the MMST with one exception: the strategy of using imagery, $F(1, 71) = 4.86$, $p = .031$. Perceived degrees of the overall and individual aspects of strategic competence involved in the MMST across the assigned task scores (low, moderate, and high) generally showed a relatively upward pattern except the strategy of using imagery that showed similar degrees across the assigned score groups and the strategy of evaluating the content that showed a slightly downward pattern. The perceived degrees across the assigned task scores did not differ between the ILST and the MMST.

To sum up, students who took either the ILST or the MMST perceived a similar degree of interactional authenticity regarding language knowledge for both assessment tasks. Similar degrees of perception for the two tasks were also found across the assigned task scores. For strategic competence, students in both groups perceived a similar degree of interactional authenticity, except for the strategy of using imagery that students in the MMST group perceived as higher ($M = 4.24$, $SD = 1.62$) than those in the ILST group ($M = 3.33$, $SD = 1.90$). It appears that content visuals included in the MMST encouraged students to use imagery more frequently for understanding and remembering information given by the input lecture. However, perceived degrees of interactional authenticity regarding strategic competence across the assigned task scores were similar for the two assessment tasks. Generally, students indicated that they perceived both the ILST and the MMST performed

similarly in tapping students' language ability consisting of language knowledge and strategic competence.

The third research question concerned elicited language knowledge manifested in student spoken responses to the ILST and the MMST. Three sub-research questions were raised to address different aspects of language knowledge. The first sub-research question investigated the use of processes, in other words, the meaning of the verbs in clauses, in the ILST and the MMST, and also across the three assigned task scores. Students in both the ILST and the MMST group, on average, produced a total of eight processes in their one-minute responses (M = 7.96, Mdn = 8.00 for the IMST; M = 8.19, Mdn = 8.00 for the MMST). Among the seven types of processes, identifying relational processes (e.g., "to be" used for giving the entity in question a definite identity) were used most by students in both groups (M = 2.96, Mdn = 3.00 for the IMST; M = 3.38, Mdn = 3.00 for the MMST). The finding met the expectation because this process is one of the key process types in the assessment tasks that require explaining two aspects of a concept. The means and medians for both groups were over two, indicating that the majority of students managed to successfully use identifying relational processes at least twice or more to achieve the target communicative goal. The groups did not significantly differ regarding the association between the type of assessment tasks and the use of identifying relational processes $\chi^2(1) = 0.79$, $p < .373$. The student use of this process type did not significantly differ in the ILST and the MMST depending on the assigned task score, $\chi^2(2) = 0.73$, $p < .695$.

Another key process type, existential processes (e.g., "there are" used for stating something exists), were not used as often by students in the ILST (M = 0.39, Mdn = 0.00), whereas students in the MMST group used this process once, on average, in their spoken

responses ($M = 0.64$, $Mdn = 1.00$). The association between the types of assessment tasks and the use of existential processes was statistically significant, $\chi^2(1) = 4.86$, $p < .027$. Based on the odds ratio, the odds of students using existential processes were 2.56 times higher if they took the MMST than the ILST. The content visuals included in the MMST input lectures possibly assisted students' understanding and recall of two aspects being taught, and choice of an existential process for including such information in their spoken responses. However, the use of this process for the two groups did not significantly differ depending on the assigned task score $\chi^2(2) = 1.01$, $p < .604$.

The second sub-research question examined the use of processes between the two task versions, biology and business, of the ILST and the MMST. The biology lecture in the ILST and the MMST contained many instances of material processes (48.8%) such as "to sharpen" and "to pull" to describe the physical actions of animals used as supporting examples in the lecture. Students who took this version of the ILST and the MMST also used material processes frequently, on average, approximately four times out of eight to nine processes in total, when they talked about the animal examples taught in the input lecture. In the business lecture of the ILST and the MMST, mental processes such as "to determine" and "to describe" (19.2%) and attributive relational processes such as "to be" (15.4%) were often used to refer to thinking processes involved in consumer decision-making and the features of factors to consider. A similar pattern was found in student spoken responses to the business version of the ILST and the MMST. Students in both groups used mental and attributive relational processes each, on average, once out of a total of eight processes. These findings suggest that the types of language in input lectures, determined by their topics, can be shown

to influence language choices students make in their spoken responses if relevant aspects of language choices are examined.

The last sub-research question focused on the use of pronouns as exophoric reference (e.g., "we/our/us" referring to the speaker and the imagined audience) between the ILST and the MMST, and across the three assigned task scores. On average, students in the ILST group used approximately one exophoric pronoun in their spoken responses, $M = 1.09$, $Mdn = 0.00$, and *range* = 0 – 10, whereas those in the MMST group used this linguistic feature slightly more than once, $M = 1.30$, $Mdn = 1.00$, and *range* = 0 – 7. There was a statistically significant association between the type of assessment tasks and the use of pronouns as exophoric reference $\chi^2(1) = 5.68$, $p < .017$. Based on the odds ratio, the odds of students using pronouns as exophoric reference were 2.75 times higher for the MMST than the ILST. It seems that visual information in the MMST assisted students in having a better sense of the audience and using exophoric pronouns to address those imagined listeners. However, the use of this linguistic feature between the two groups did not significantly differ depending on the assigned task score $\chi^2(2) = 1.99$, $p < .369$, indicating that the use of exophoric pronouns across the assigned task scores did not differ between the ILST and the MMST.

All in all, both the ILST and the MMST functioned well in eliciting the use of the following features of the target language knowledge: identifying relational processes, material processes for the biology version, and mental and attributive relational processes for the business version. However, existential processes and pronouns as exophoric reference were elicited more successfully by the MMST than by the ILST. According to these linguistic analyses of student spoken responses, the extent of involvement of examinees' language knowledge required to accomplish an assessment task, in other words, interactional

authenticity regarding language knowledge, was higher for the MMST. However, the lack of association between these features and raters' judgments about overall level of performance indicate that these features may not have been judged as important for distinguishing among examinees' performance.

The fourth research question concerned strategic competence involved in accomplishing the ILST and the MMST. In this study, strategic competence was operationalized as *reported* actions and thought processes of the examinees during stimulated recall interviews, and categorized following Swain et al.'s (2009) strategy scheme. Ten students who took the ILST and another ten who took the MMST from the Spring 2013 administration participated in stimulated recall interviews.

Students in both the ILST and the MMST group reported a frequent use of communication, cognitive, and metacognitive strategies. The mean reported counts of communication and metacognitive strategies were relatively similar between the two groups. However, students in the MMST group reported a more frequent use of cognitive strategies (M = 7.90, Mdn = 8.00, and *range* = 2 – 13) than those in the ILST group (M = 4.90, Mdn = 5.00, and *range* = 2 – 8). Specifically, the MMST students reported more instances of the strategies of anticipating the content (M = 3.50, Mdn = 3.00, and *range* = 0 – 6) and using imagery (M = 1.40, Mdn = 1.00, and *range* = 0 – 4) than the ILST students (M = 1.80, Mdn = 1.50, and *range* = 1 – 4 for the strategy of anticipating the content; M = 0.50, Mdn = 0.00, and *range* = 0 – 2 for the strategy of using imagery). Often, these two frequent strategy uses were facilitated by content visuals included in the MMST. Additional visual channels appeared to encourage students to use the strategies of anticipating the content and using imagery for good comprehension and recall. Across the assigned task scores (low, moderate,

and high), the mean reported counts generally showed an upward pattern. However, students in the MMST group obtaining low and moderate task scores, on average, reported high mean counts of the cognitive strategies (M = 8.00 for low; M = 7.83 for moderate) similar to those of the MMST students obtaining high task scores (M = 8.00). It seems that students in the MMST group used this strategy type regardless of their English proficiency and content visuals in the input lectures may have facilitated a frequent use for all students.

5.2. Implications of the Study

The findings of the current study have theoretical, practical, and methodological implications. Theoretical implications concern the authenticity of integrated assessment tasks, specifically integrated listening-speaking tasks. Practical implications include the role of visuals in integrated tasks and the design of a task-specific, language-based rating scale descriptor for evaluating the content of spoken responses. Methodological implications include 1) a comprehensive analysis of elicited language ability based on the interactionalist perspective and 2) a new discourse analytic tool used for studying language samples from second language assessments. The following sub-sections discuss each implication.

5.2.1. Theoretical Implications

Bachman (1991) proposed an approach to characterizing authenticity as two types, 1) situational and 2) interactional authenticity. The current study adopted this approach and managed to investigate the authenticity of the ILST and the MMST from various angles. Language testing researchers and test developers have long preferred integrated assessment tasks, the assessment task type of the ILST and the MMST, due to their presumed authentic

nature (Plakans, 2013; Weir, 1990). However, such authenticity has been taken for granted without empirical justification. It has been unknown 1) how relevant the characteristics of integrated assessment tasks are to those of the target language use (TLU) tasks (*situational authenticity*), and 2) to what extent language knowledge, processes, and strategic competence involved in accomplishing integrated assessment tasks, correspond to those in the TLU tasks (*interactional authenticity*) (Hulstijn, Coopmans, van Hout & Bos, 2003).

The current study is the first to provide empirical evidence to evaluate the apparent authenticity of integrated assessment tasks, specifically integrated listening-speaking tasks[13]. The findings suggested that students, who are in one of stakeholder groups in the TLU situation, perceived a relatively high degree of situational authenticity of this task type overall.

However, students' perceptions of interactional authenticity regarding their use of specific aspects of language knowledge varied. The perceived degrees of using knowledge of mental, relational, and existential processes were consistent with those for expected language knowledge use in the TLU task, whereas the perceived degrees of using knowledge of material processes and exophoric pronouns were lower. Based on the linguistic analyses of what students actually produced in their spoken responses, use of material, mental, and relational processes corresponded to what is expected in the TLU task; on the other hand, use of existential processes and exophoric pronouns was less than expected. In summary, the interactional authenticity regarding language knowledge of integrated listening-speaking tasks was high in some regards, but not in others.

[13] The authenticity of the multimedia-mediated speaking task will be discussed in Sub-section 5.2.2 where practical implications are described.

Students' perceptions of interactional authenticity regarding strategic competence also varied. The perceived degrees of using the strategies of anticipating the content, using imagery, planning, and evaluating language production were only moderate. Based on student reports during stimulated recall interviews, the strategies of using imagery, evaluating language production, referring to notes, and setting goals were all moderately involved. Overall, the interactional authenticity regarding strategic competence of integrated listening-speaking tasks was moderate.

In summary, the popular belief concerning the authenticity of integrated tasks, specifically integrated listening-speaking tasks, was partially supported by the findings of the current study. The situational authenticity of the tasks was perceived as high overall; however, varying degrees of different aspects of interactional authenticity were found. A few features of the target language knowledge were not recognized by students as being sufficiently elicited or not manifested enough in their spoken responses, and the strategic competence defined in the task construct was reported as only moderately involved overall during task accomplishment. The performance from these tasks may not be representative speech samples from which an inference of an individual's language ability in the TLU situation can be drawn, and need to be interpreted with caution. Integrated tasks in academic language proficiency tests have been regarded as authentic because they are seen as capable of assessing a kind of language ability crucial in an academic context. However, the elicited language performance may not fully capture the target language ability, contrary to the common belief.

5.2.2. Practical Implications

The current study is noteworthy in that the multimedia-mediated speaking task (MMST) I developed is the first integrated task that incorporates visuals and manages to create a context where the task can elicit a large amount of target language performance from examinees. This innovative task elicited more features of the target language knowledge than did the integrated listening-speaking task (ILST). In addition, the MMST fostered students' use of imagery to understand, think, or remember information. Together, students who took the MMST may have executed the strategy of using visuals more frequently to understand the task context and made appropriate language choices in response to the identified contextual features.

This suggests the reasonable possibility of incorporating visuals in developing integrated assessment tasks. First of all, content visuals added to the input materials may have potential for assisting comprehension and recall and leading to accurate representations of meaning. In the case of the MMST featured in this study, students who took this type of integrated task reported more frequent use of imagery as a strategy to understand and remember information taught in the input lectures, and had a more frequent use of existential processes to state the classification of a concept (e.g., **there are** two competing definitions of tools). The addition of visuals may have helped comprehension, as also found in some previous studies on second language listening tests using video materials (e.g., Ginther, 2002; Sueyoshi & Hardison, 2005; Wagner, 2006, 2010b), and provided opportunities to linguistically express the comprehended information. This suggests that examinees in such a condition are less challenged regarding their understanding of a fairly complex concept learned in a few minutes, and therefore, not limited in their attempts to make meanings of

key information. This can perhaps reduce the uncertainty of variability sources, comprehension or production, in integrated tasks. Conceptually, adding visuals to an integrated task may expand the task context which allows various semiotic systems to be used, and enables "oral communication" that may closely resemble real-life. This enhancement may overcome the limitation of the current audio-only integrated tasks that simply combine two skills, "listening" and "speaking," with insufficient contextual information.

Secondly, additional visual elements in integrated tasks, such as a professor moving in the input video, can possibly simulate a setting close to the TLU situation more successfully, and provide a sense of the imagined audience over the other side of the computer screen to which examinees speak during the task. This may encourage examinees to select language with audience in mind. In the current study, students who took the MMST had a more frequent use of exophoric pronouns such as "we" and "you" than those who took the ILST to address the audience who would potentially listen to their responses. Some people criticize one-way communication in most of the current computer-based speaking tests because it elicits samples of monologue that is not particularly constructed for transmitting information to counterparts. This type of spoken sample allows only a limited scope of inferences about an individual's language ability. However, if the adding of visuals into integrated tasks can overcome the context of talking to a "machine" by facilitating an awareness of potential listeners on the other side, we can still maintain the efficiency of computer-based tests and at the same time, obtain speech samples more relevant to a real-life situation.

In addition, the findings from the linguistic analyses suggest to test developers the design of a task-specific rating scale descriptor for content of speech samples. It was found in the current study that the topics of input lectures and the types of language used in the lectures influenced the language choices in examinees' responses. Frost et al. (2012) used an intuitively constructed list of key points of the input texts in validating part of their content-related ratings scale. On the other hand, ideational comparison between input texts and spoken responses will possibly provide a more language-based measure to evaluate the expression of content. Such a measure has the potential for capturing examinees' performance more systematically in terms of topic development and corresponding language choices.

5.2.3. Methodological Implications

This study contributes methodologically to the language-testing field in that it provides a comprehensive analysis of elicited language ability based on the interactionalist perspective to construct definition. The study examined language knowledge, strategic competence, and their relationships to the context of a task altogether. The conceptual framework of the interactionalist approach has provided an explicit approach for conceptualizing language ability with implications for test development and validation for almost two decades (Chapelle, 1998; Read & Chapelle, 2001). However, empirical studies using this framework tend to investigate each component of language ability separately, mostly on strategic competence (Phakhiti, 2003; 2008), and not consider the whole.

The comprehensive analysis of elicited language ability in this study allowed interpretations of dynamics of language performance as a collective unit. The findings of

both linguistic and strategic analyses on the same assessment tasks taken by the same sample of examinees provide a larger perspective on how language knowledge and strategic competence in response to the contextual features of the task affect task performance. For instance, in the current study, students who took the MMST used existential processes, one of the target language features expected for successful task performance, more frequently, and reported more frequent use of imagery as a strategy to understand and remember information than those who took the ILST. These findings made it possible to offer an interpretation that the strategy of using imagery helped in comprehending and recollecting the content of the input lecture and this cognitive processing tapped examinee knowledge of language expected in the task context and using it to express meanings of understood and remembered information.

Additionally, the linguistic analyses of specific systemic functional linguistic (SFL) features in the current study add to a repertoire of analytic tools used for studying language samples from second language assessments. Although research on language knowledge interacting with context exists in the setting of formative classroom assessment (Leung & Mohan, 2004; Low, 2010; Mohan, Leung, & Slater, 2010; Slater & Mohan, 2010), the current study is original in using this approach to analyze language knowledge elicited by the tasks designed for large-scale assessment. The analysis of specific language features expected in this context allowed an interpretation and explanation of examinees' language choices influenced by contextual factors. This approach is conceptually in line with the view of language knowledge in an interactionalist perspective to construct definition, and methodologically provides a system to investigate meaning-based language choices that test takers make.

5.3. Limitations of the Study

The findings of the current study, admittedly, need to be interpreted with caution as they are not free from limitations. First, the ratings on the two authenticity questionnaires, situational and interactional, were based on student perception and resulted from a reflection of each individual's interpretations on the items being asked. Individuals may have perceived a magnitude of intervals on the six-point Likert scale, on which the authenticity questionnaires were based, as different, and therefore responses of the same rating may have actually had different meanings. In addition, since student ratings of the interactional authenticity questionnaire were based on what students thought they had done a few minutes earlier during an assigned assessment task, their memory may have faded, or such cognitive or metacognitive processing may have not been consciously noticed and/or recalled. In this sense, the questionnaire ratings may have captured only a partial view of how students actually executed their language ability. In fact, there were mismatches between student perception of use of their language ability and actual use manifested in student production. This suggests that data based on perception should be interpreted cautiously.

Second, the equivalent characteristics between the ILST and the MMST groups in terms of overall academic English proficiency were assumed only on the basis of stratified-random assignment. Students in the current study were recruited from four different sources each of which can be interpreted as either a low, moderate, or high English proficiency level. Students were originally placed into these groups according their TOEFL scores. For example, Group 1 consisted of students who were conditionally admitted to the university with a TOEFL total score of 55 or higher, whereas Group 4 consisted of those who were admitted to the university with a TOEFL admissions requirement total score of 111. Using

these group-level criteria for characterizing individuals' overall proficiency was practical, but may have resulted in some inaccurate group assignments for assessing performance in the integrated assessment tasks investigated in the current study. Using proficiency measures of an individual's ability of utilizing oral communication skills, when checking group equivalence, will ensure a ground for more valid comparative analyses.

Third, some of the analyses conducted in this research were based on a relatively small number of study participants in each analytic unit; thus, the results from those analyses may have been sample-dependent. Although adequate numbers were in the two main units for the study—46 students who took the ILST and 47 who took the MMST, there were only a few students placed in one analytic unit (six the minimum) for the analyses across the three assigned task scores (low, moderate, and high). In addition, the analysis of strategic competence was based on the stimulated recall data from only 20 students (ten for each task type), and the interpretations of quantitatively transformed strategic behavior reports were based solely on the descriptive statistics. To draw more generalizable conclusions, data from a larger number of participants are suggested. The results from a broader pool of samples would expand the scope of interpretations I can make, and support (or oppose) the findings from the current study with more convincing evidence.

5.4. Suggestions for Future Research

The findings of the current study suggest ideas for modification and expansion and directions for future research. First, the situational authenticity questionnaire (Appendix G) with more sensitive questionnaire items and rating scale may be able to measure possible variance at the upper end of the current scale and provide more informative data for

understanding the perceived degree of situational authenticity. The distributions of the current data sets on most of the situational authenticity questionnaire items were left-skewed, reflecting a ceiling effect in participant ratings. A more sensitive questionnaire may capture an accurate picture of participant perception and in addition, measure possibly subtle differences in the degree of situational authenticity between the ILST and the MMST.

Second, the set of analyses, conducted in the current study, for other types of integrated tasks such as ones that combine reading, listening, and speaking skills will expand the understanding about the function of integrated tasks overall. The current study investigated how an integrated listening-speaking task and its modified type, a multimedia-mediated speaking task, 1) simulate a situation corresponding to the target language use (TLU) situation and 2) elicit samples of examinees' performance in using the target language ability. The analyses of these aspects for other types of integrated tasks will have implications for task development such as establishing informed guidelines for selecting and incorporating different task features, and validation such as providing evidence of representativeness and relevance of the assessment tasks.

Third, other stakeholder groups in an academic domain (e.g., professors) can also be considered in examining the situational authenticity of integrated tasks. In the current study, students were asked to indicate their perception of how the features of the ILST and the MMST corresponded to those of the TLU task. However, professors are the other group of stakeholders in the TLU situation of interest, and investigating their perception also should provide a fuller understanding of the situational authenticity from both ends of the communication, although practical challenges of participant recruitment should be noted.

Fourth, the relationship between the types of language used in the input lectures and language choices made by examinees in their spoken responses can be investigated further with various other topics. The current study found the influence of lecture topics, biology and business, on types of processes examinees chose to use. It will be interesting to see if this association is found with other different lecture topics in biology and business and also in different academic disciplines. Findings from these kinds of follow-up studies together with those from the current study will help us to understand the specific linguistic relationship between test tasks and the actual responses students produce in integrated tasks, and provide useful information in task development, particularly in designing a task-specific rating scale descriptor for content of speech samples. Specifying the main language types of the input lectures in the content aspect of rating scale descriptors can assist raters to pay attention to these linguistic features in examinee spoken responses and systematically evaluate language examinees used to explain the main content they learned from the input lectures.

Lastly, another suggestion for future research is to consider the effects of nonverbal communication features in the video input of the MMST on student task performance. The current study focused on the content visuals included in the input lectures as the main contributing factor of possible difference in student performance between the ILST and the MMST. However, the input videos also included other nonverbal communication features such as sound of voice, gestures, and gender of the speaker. Previous research suggests that these other features affect the manner of communication (Moore, Hickson, & Stacks, 2009). Studies examining the effects of different nonverbal communication factors will increase understanding of how additional visual elements interact with examinees and possibly affect students' use of language.

5.5. Conclusion

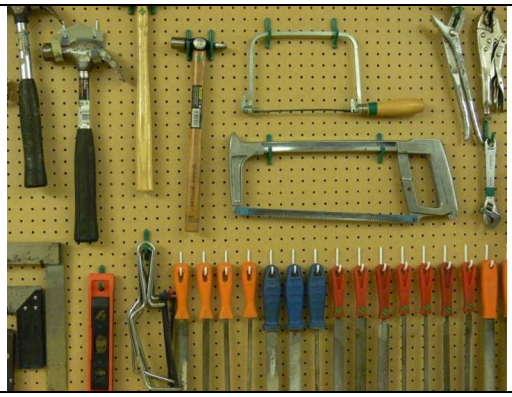
This dissertation introduced an authenticity analysis for an existing integrated task type, an integrated listening-speaking task (ILST), and a newly developed type, a multimedia-mediated speaking task (MMST). The findings from this analysis can provide backing/rebuttal for the domain definition inference of a validity argument for an academic English proficiency test that contains the task types similar to the ILST and/or the MMST. Some support and others do not support the task modeling assumption, meaning the types of integrated tasks featured in the current study can simulate tasks that are representative of the academic domain and require important language ability. Specifically, findings that support the representativeness of the integrated tasks investigated in the current study are student ratings on the situational authenticity questionnaire. Those that support the relevance of elicited language performance are student ratings on the interactional authenticity questionnaire and the MMST student spoken responses and strategic behavior reports. Findings that fail to support the relevance of elicited language performance are the ILST student spoken responses and strategic behavior reports.

Overall, this dissertation should provide an understanding of the nature of the innovative task type, integrated assessment tasks, and better inform task selection in the course of test development. Further, it contributes to interpreting the meaning of performance from this type of task, and eventually, investigating the validity of interpretations and uses of an academic English proficiency test that contains the similar types of integrated tasks.

APPENDIX A

CONTENT VISUALS FOR THE MULTIMEDIA-MEDIATED SPEAKING TASK (MMST)

Task Version A - Biology

<table>
<tr><th>Slide 1</th><th>Slide 2</th></tr>
<tr><td>Illustrate the oral stimulus</td><td>Organize information in the stimulus</td></tr>
<tr><td>  </td><td>Definition of Tool

1.Narrow definition
2.Broad definition</td></tr>
<tr><th>Slide 3</th><th>Slide 4</th></tr>
<tr><td>Replicate the oral stimulus</td><td>Illustrate the oral stimulus</td></tr>
<tr><td>Narrow Definition

•A purposefully changed or shaped object</td><td>  </td></tr>
<tr><th>Slide 5</th><th>Slide 6</th></tr>
<tr><td>Replicate the oral stimulus</td><td>Illustrate the oral stimulus</td></tr>
<tr><td>Broad Definition

•Any object used to perform a specific task</td><td>  </td></tr>
</table>

Task Version B - Business

Slide 1	**Slide 2**
Illustrate the oral stimulus	Organize information in the stimulus
	Factors of Product Quality 1. Reliability 2. Features
Slide 3	**Slide 4**
Replicate the oral stimulus	Illustrate the oral stimulus
Reliability • The absence of unexpected defects or problems	
Slide 5	**Slide 6**
Replicate the oral stimulus	Illustrate the oral stimulus
Features • Extras that make a product easier to use or attractive	

APPENDIX B

DESCRIPTION OF TASK FORMAT FOR THE MMST

Slide 1:
Visual: a student wearing a headset with a written statement saying "Please watch carefully."
Audio: "In this question, you will watch a short lecture. You will then be asked to explain important information from the lecture. After you hear the question, you will have 20 seconds to prepare your response and 60 seconds to speak."

Slide 2:
Visual: a professor next to a screen with a PowerPoint slide
Audio: "Now watch part of a lecture in a biology (task version A)/ business (task version B) class."

Slide 3:
Video: a biology (task version A)/ business (task version B)] lecture
(The participants watch an approximately two-minute lecture.)

Slide 4:
Visual: the written instructions below
Audio: "Now get ready to answer the question. You may use your notes to help you answer."

Slide 5:
Visual: the written prompt below
Audio:

Task version A: "Using points and examples from the lecture, describe the two different definitions of tools given by the professor."
Task version B: "Using points and examples from the lecture, explain the two major factors of product quality and how their role in consumer decision making has changed."

Slide 6:
Visual: the written information below:

Preparation time: 20 seconds
Response time: 60 seconds,
and a digital stopwatch counting 20 seconds

Audio: "Begin to prepare your response after the beep."
(The participants have 20 seconds to prepare their response.)

Slide 7:
Visual: numeric countdown from 60 seconds
Audio: "Begin speaking after the beep."
(*The participants have 60 seconds to speak.*)

Slide 8:
Visual: the following written sentence, “This is the end of the task.”

APPENDIX C

TASK SPECIFICATIONS FOR THE MMST

I. Definition of the ability to be assessed

An ability to appropriately and intelligibly convey key ideas from a short lecture segment on academic topics

II. Characteristics of the setting in which the tasks are administered

A. Physical characteristics

1. **Conference room on campus, quiet, comfortable**
2. **Equipment**
   a) Each examinee provided with a laptop computer
   b) External microphone for recording
   c) Degree of familiarity: all quite familiar with computers and recording process

B. Participants

1. **Examinees**: undergraduate and graduate students who are non-native speakers of English
2. **Administrator**: the researcher experienced in using the equipment and having a positive attitude toward the examinees

C. Time of task: by appointment based on the examinees' schedule

III. Characteristics of the input, expected response, and relationship between input and expected response

A. Input

1. **Format**
   a) Channel: aural and visual (a video lecture on the computer)
   b) Form: language and non-language (context and content visuals)
   c) Language: English (target)
   d) Length: extended discourse (300 – 350 words approximately in two minutes)
   e) Vehicle: reproduced
   f) Degree of speededness: normal (approximately 160 words per minute)
   g) Type: input for interpretation (videotext)
2. **Language of input**
   a) Language characteristics
      Organizational characteristics: as occurs in lectures from university courses
      (a) Grammatical
      (i) Morphology and syntax: wide range of organized structures
      (ii) Vocabulary: wide range of general and technical vocabulary
      (iii) Phonology: standard American English
      (b) Textual (cohesion and organization): wide range of cohesive devices and rhetorical organizational patterns, including narration, description, definition, classification, and comparison and contrast
   b) Pragmatic characteristics: as occurs in lectures from university courses

(1) Functional: ideational and heuristic
(2) Sociolinguistic: standard dialect, formal/informal register, natural

3. **Topical characteristics**: academic

B. Expected response

1. **Format**
   a) Channel: oral
   b) Form: language
   c) Language: English (target)
   d) Length: extended discourse in 60 seconds
   e) Degree of speededness: normal
   f) Type: extended production response
2. **Language characteristics**
   a) Organizational characteristics: vocabulary similar to that in the video lecture; morphology and syntax: standard English; phonology: standard American English
   b) Pragmatic characteristics: mostly same as the video lecture, plus some need for appropriate register use
3. **Topical characteristics**: academic

C. Relationship between input and expected response

1. **Type of external interaction**: Interrelatedness: non-reciprocal
2. **Scope of relationship**: both narrow and broad—narrow because specific pieces of information must be provided, and broad because the relationship between those chosen information and the whole video lecture must be kept in mind by examinees
3. **Directness of relationship**: direct

IV. Instructions for responding to task

Directions: Now get ready to answer the question. You may use your notes to help you answer. "Using points and examples from the lecture, describe/explain the two (main points of the video lecture)." Begin to prepare your response after the beep. (After 20 seconds) Begin speaking after the beep.

Updated from the framework of Bachman & Palmer (2010), pp. 318 - 320

APPENDIX D

DESCRIPTION OF TASK FORMAT FOR THE INTEGRATED LISTENING-SPEAKING TASK (ILST)

Slide 1:
Visual: a student wearing a headset with a written statement saying "Please listen carefully."
Audio: "In this question, you will listen to a short lecture. You will then be asked to explain important information from the lecture. After you hear the question, you will have 20 seconds to prepare your response and 60 seconds to speak."

Slide 2:
Visual: a professor in front of a chalkboard
Audio: "Now listen to part of a lecture in a biology (task version A)/ business (task version B) class."

Slide 3:
Visual: the professor from Slide 2
Audio: a biology (task version A)/ business (task version B) lecture
(The participants listen to an approximately two-minute lecture.)

Slide 4:
Visual: the written instructions below
Audio: "Now get ready to answer the question. You may use your notes to help you answer."

Slide 5:
Visual: the written prompt below
Audio:

> Task version A: "Using points and examples from the lecture, describe the two different definitions of tools given by the professor."
> Task version B: "Using points and examples from the lecture, explain the two major factors of product quality and how their role in consumer decision making has changed."

Slide 6:
Visual: the written information below:

> Preparation time: 20 seconds
> Response time: 60 seconds,
> and a digital stopwatch counting 20 seconds

Audio: "Begin to prepare your response after the beep."
(The participants have 20 seconds to prepare their response.)

Slide 7:
Visual: numeric countdown from 60 seconds

Audio: “Begin speaking after the beep.”
(*The participants have 60 seconds to speak.*)

Slide 8:
Visual: the following written sentence, “This is the end of the task.”

APPENDIX E

TASK SPECIFICATIONS FOR THE ILST

I. Definition of the ability to be assessed

An ability to appropriately and intelligibly convey key ideas from a short lecture segment on academic topics

II. Characteristics of the setting in which the tasks will be administered

A. Physical characteristics

1. **Conference room on campus, quiet, comfortable**
2. **Equipment**
   a) Each examinee provided with a laptop computer
   b) External microphone for recording
   c) Degree of familiarity: all quite familiar with computers and recording process

B. Participants

1. **Examinees**: undergraduate and graduate students who are non-native speakers of English
2. **Administrator**: the researcher experienced in using the equipment and having a positive attitude toward the examinees

C. Time of task: by appointment based on the examinees' schedule

III. Characteristics of the input, expected response, and relationship between input and expected response

A. Input

1. **Format**
   a) Channel: aural
   b) Form: language and non-language (context visuals in still pictures)
   c) Language: English (target)
   d) Length: extended discourse (300 – 350 words approximately in two minutes)
   e) Vehicle: reproduced
   f) Degree of speededness: normal (approximately 160 words per minute)
   g) Type: input for interpretation (audio text)
2. **Language of input**
   a) Language characteristics
      Organizational characteristics: as occurs in lectures from university courses
      (a) Grammatical
      (i) Morphology and syntax: wide range of organized structures
      (ii) Vocabulary: wide range of general and technical vocabulary
      (iii) Phonology: standard American English
      (b) Textual (cohesion and organization): wide range of cohesive devices and rhetorical organizational patterns, including narration, description, definition, classification, and comparison and contrast
   b) Pragmatic characteristics: as occurs in lectures from university courses

(1) Functional: ideational and heuristic
(2) Sociolinguistic: standard dialect, formal/informal register, natural

3. **Topical characteristics**: academic

B. Expected response

1. **Format**
   a) Channel: oral
   b) Form: language
   c) Language: English (target)
   d) Length: extended discourse in 60 seconds
   e) Degree of speededness: normal
   f) Type: extended production response
2. **Language characteristics**
   a) Organizational characteristics: vocabulary similar to that in the audio lecture; morphology and syntax: standard English; phonology: standard American English
   b) Pragmatic characteristics: mostly same as the audio lecture, plus some need for appropriate register use
3. **Topical characteristics**: academic

C. Relationship between input and expected response

1. **Type of external interaction**: Interrelatedness: non-reciprocal
2. **Scope of relationship**: both narrow and broad— narrow because specific pieces of information must be provided, and broad because the relationship between those chosen information and the whole audio lecture must be kept in mind by examinees
3. **Directness of relationship**: direct

IV. Instructions for responding to task

Directions: Now get ready to answer the question. You may use your notes to help you answer. "Using points and examples from the lecture, describe/explain the two (main points of the audio lecture)." Begin to prepare your response after the beep. (After 20 seconds) Begin speaking after the beep.

Used the framework of Bachman & Palmer (2010), pp. 318 – 320

APPENDIX F

RESEARCH PROTOCOL OF STIMULATED RECALL

The researcher

- Operates a built-in computer camera and makes notes during the participant's task performance.
- Replays the participant's videotaped performance during the stimulated recall session.
- Carries out stimulated recall (selecting segments of the video clip to examine).
- Ensures audio recording is working during stimulated recall.

Stimulated Recall Instructions:

1. Provide explanation of the stimulated recall session:

"What we're going to do now is watch the video. I am interested in what you were thinking at the time you were doing the task. I can see what you were doing by looking at the video, but I don't know what you were thinking. So what I'd like you to do is tell me what you were thinking, what was in your mind at that time while you were doing the task, not what you think about it now.

You can pause the video any time that you want using the touchpad or simply pressing the space bar. So if you want to tell me something about what you were thinking, you can push pause. If I have a question about what you were thinking, then I will push pause and ask you to talk about that part of the video. You can stop the video as often as you want, and we don't need to take turns."

2. Demonstrate stopping the video, and have the participant try using pause button.
3. Ask if there are questions about procedure:

"Is it clear what we're doing? I want to know what you were thinking as you did the task."

4. Begin video.
5. If the participant stops the video, listen to what he or she says.

If the participant starts to talk without pausing video, pause video for him/her, and then wait for him/her to unpause it. If he or she doesn't, wait for a few seconds of silence, and then ask, "Do you remember anything else about what you were thinking at that moment?"

If the researcher stops the video, ask something general, for example:
"What were you thinking here/ at this point/ right then?
Can you tell me what you were thinking at that point?
I see you're laughing/ looking confused/ saying something there, etc. What were you thinking then?"

6. If the participant's response is that he or she doesn't remember, do not pursue this because "finishing" for answers that were not immediately provided increases the

likelihood that the answer will be based on what the person thinks now or some other memory or perception.

7. Try not to focus or direct participant responses beyond "what were you thinking then."
8. Try not to react to responses other than providing backchanelling cues or nonresponeses:

"Oh, mhm, great, good, I see, uh-hum, ok"

9. When the participant has *finished* the recall, ask explicit questions, if necessary, to elicit more data to address the research question, how strategic competence is involved to complete the assessment tasks. Some example questions may be: what did you first think right after you were given the task?; did you examine the characteristics of the assessment task to determine what resources are needed to complete it?; did you select elements from the areas of topical and language knowledge for successfully completing the assessment task?
10. Ask the participant if he or she has any questions or comments about the video or the task he or she has done.

adapted from Gass & MacKey (2000), pp. 153-160

APPENDIX G

SITUATIONAL AUTHENTICITY QUESTIONNAIRE

Directions: Compare the task you did with a task you would do in a real-life academic context. Please circle a number to indicate your opinion and explain your answer.	*Very different from real-life academic task*					*Very close to real-life academic task*
1. Do you understand the purpose and procedures of doing the task based on the information given?	1	2	3	4	5	6
2. As you were doing the task, was there enough information about the location and how it looked to picture yourself in the real-life academic task there?	1	2	3	4	5	6
3. Does the person speaking in the lecture seem like a person you might encounter in a real-life academic context?	1	2	3	4	5	6
4. Do you think there could be a purpose of doing the task in a real-life academic context, apart from assessing your English ability?	1	2	3	4	5	6
5. Would you encounter this type of task format in a real-life academic context?	1	2	3	4	5	6
6. Do you hear the pronunciation, vocabulary, grammar, cohesion and organization used in the lecture in the real-life academic task?	1	2	3	4	5	6
7. Would you use English used in your spoken response in the real-life academic task?	1	2	3	4	5	6
8. Is the problem you had to solve in responding to the task prompt one that you might have to solve in a real-life academic context?	1	2	3	4	5	6
9. Is the relationship between the lecture and your spoken response one that you would encounter in a real-life academic context?	1	2	3	4	5	6
10. Is this type of task complete and realistic?	1	2	3	4	5	6

adapted from Liu (2005), p. 123

APPENDIX H

INTERACTIONAL AUTHENTICITY QUESTIONNAIRE

Directions: Consider how much the following aspects were involved in accomplishing the task you just finished. Please circle a number to indicate your opinion and explain your answer.	*Not at all*					*A lot*
1. Knowledge of language to give the entity a definite identity (e.g. (*the entity*) is (*a definite identity*); (*the entity*) means (*a definite identity*))	1	2	3	4	5	6
2. Knowledge of language to ascribe some descriptive attributes to an entity (e.g. (*an entity*) is (*a descriptive attribute*))	1	2	3	4	5	6
3. Knowledge of language to give a sense of physical action	1	2	3	4	5	6
4. Knowledge of language to realize the mind and senses	1	2	3	4	5	6
5. Knowledge of language to state that something exists (e.g. there are…)	1	2	3	4	5	6
6. Knowledge of language to refer to someone outside the context of your spoken response (e.g. we, you, your…)	1	2	3	4	5	6
7. Strategy of seeking a goal for completing the task	1	2	3	4	5	6
8. Strategy of anticipating the content	1	2	3	4	5	6
9. Strategy of using visual images to understand, think, or remember information	1	2	3	4	5	6
10. Strategy of using note-taking to remember or organize information	1	2	3	4	5	6
11. Strategy of reviewing the notes to remember/formulate what to say	1	2	3	4	5	6
12. Strategy of planning the parts, sequence, or main ideas to be expressed orally	1	2	3	4	5	6
13. Strategy of referring to the notes to remember/formulate what to say	1	2	3	4	5	6
14. Strategy of evaluating your language production	1	2	3	4	5	6

APPENDIX I

CODING SCHEME FOR STRATEGIC COMPETENCE ANALYSIS

	Definition/ substrategy	Example(s)
Approach strategies: What the test-taker does to orient him- or herself to the task		
Recalling the task type	Test-taker thinking about the task's format	I listen to the questions (instruction) and I know the tasks of the this lecture and which is two minutes lecture and I have twenty minutes (seconds*) to prepare for question after the lecture. Yeah, I got the information and prepare to listen to the content. (Participant 2.3)
Recalling the question	Test-taker thinking about the meaning of the question	And I think the ques- the ques- okay, the question asked to use specific examples and the points, its points and examples. (Participant 2.3)
Recalling the lecture	Test-taker thinking about the lecture	I tried to remember as much as possible what I heard from the lecture. (Participant 3.1)
Communication strategies: involving conscious plans for solving a linguistic problem in order to reach a communicative goal		
Paraphrasing	Test-taker restating in another form or with other words to clarify meaning	So instead of saying, clients, or I, I used the whole phrase I think like people who were going to buy things, instead of consumers, for example, right? Somebody who prepares more, they might say consumers, but because I'm doing and thinking in real time, if the word doesn't come to my mind, well, we have to paraphrase, right? (Participant 4.5)
Approximating	Test-taker using lexical or grammatical substitution	I didn't remember the "found" words, so I just replace that word as a "coincidence" bec-, so <laugh> I'm very, I'm very, what can I say, I, I'm very feel empty <laugh> while, no, what can I say, feel empty because I, I'm very funny about myself because I translate the "found" word as a "coincidence." Elephant find, found that stick coincidence, so it is very funny for me, for myself. (Participant 2.6)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Linking to prior experiences/ knowledge	Test-taker making connections between what is known or his/her previous experience and what he/she is hearing or watching	I remember video I have watched before about chimpanzee and how they use the, you know, the tools for hunting. And the woman, I think, she was living in Africa. She is British, and went to Africa, and study about chimpanzee. So my memory went to that. Yeah, so I just connect some idea and that's help* me to understand little bit. (Participant 1.4)
Reviewing notes	Test-taker reviewing the notes in order to remember/ formulate what to say	In this point, I need to organize and arranged what I'm gonna say, something like that, so I just look at my note-taking... (Participant 1.2)
Referring to notes	Test-taker referring to the notes in order to remember/ formulate what to say	And the second one, and yes, second one, just followed the notes, and just read the, read notes without constructed in a appropriate order or add some more words and the informations* to enrich my answers. I just read the note. (Participant 2.3)
Organizing thoughts	Test-taker organizing ideas while speaking	I wanted to talk about the introduction, the very first part, but I thought that the time, there was time limit, so I just wanted to, I just moved on the main focused two factors, so at the first time, at the very beginning of what I was saying, I thought that introduction was not necessary in this speaking... (Participant 1.2)
Guessing	Test-taker guessing by using linguistic or other clues	I try to catch what was the first definition because I didn't know the word, "narrow." So is it narrow or a word I don't know? So the word which is just appear my mind is "narrow." Yeah, yeah. (Participant 1.4)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Rehearsing	Test-taker mentally rehearsing what to say	… I just practice first sentence... (Participant 2.5)
Elaborating to fill time	Test-taker elaborating on points that might not be relevant to the question in order to fill the time	... and also want to talk some thing, talk something in order to let my speech, speak time longer, because from this word, I can just can say it out in 20 minute, 20 <laugh> seconds, but I need to said about 60 seconds, so I will find something to say, so just during this time, I just try my best to say some connection or some words to know use <laugh> in order to for the time. (Participant 2.5)
Cognitive strategies: involving manipulating the target language to understand and produce language		
Anticipating the content	Test-taker anticipating the content	I saw the two images on the screen for the red car and the blue cars. I was thinking that this lecture might be related with the car things. (Participant 2.1) I was thinking to myself like something they, the professor mentioned about like picking up a, a, a stick from the ground, so I was thinking maybe next example will be related to this one. And it will be maybe a, like this one is not picking up from a ground, from the ground or maybe that one will be picking up from the ground. (Participant 4.4)
Anticipating the structure	Test-taker anticipating the structure of talk during a listening/watching activity	And I just, I, I know the structure of the lecture is first describe the narrow and de- definition for that hand example and from this, from time till then, and I know the second parts will be the same. The definition of the broad way and the example. (Participant 2.5)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Anticipating the question	Test-taker anticipating the question	I little guess about they will ask about the two definition because the professor said only about tool and the two definition about the tools, so I can predicts. (Participant 2.6)
Using imagery	Test-taker using visual images, either generated or actual, to understand, think, or remember information	So in this case, the balance, the balance, the, the, the illustration of balance help me, like to help me to confirm what, what I, what I'm listening like it's not deciding factor, factor, like reliability is not deciding factor, factors, and yeah, the, the visual aids help me to, to confirm this, and like pay more attention to the second one like, the second one is more important than, than the previous one. So, yeah, so the visual aids help me more to, to confirm what I'm listening. (Participant 4.1) … in terms of the whole lecture, what I'm trying to do is picture everything he's talking about, right? And it's very concrete because you can imagine a car, you can imagine things like that, right? (Participant 4.5)
Using mechanical means to organize or remember information	Test-taker writing things down to organize or remember information	Then, I note down, I jot down the two main points, the two factors on the paper. That's help me to remember what he talk. (Participant 3.1)
	Test-taker using symbols for drawing attention during delivery	I have an upside smile here, so of course, if I listen to something that I don't agree with, I will have some sort of emotional reaction to it, right? (Participant 4.5)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Using mechanical means to organize or remember information	Test-taker writing down information in numerical form during a listening/watching activity	Okay, the number two, two definitions, so I write down "2." (Participant 2.3)
	Test-taker mapping information to organize notes during a listening/watching activity	And I kind of draw this, I can show you, like this outline, kind of cause and effect outline here that will help me to, you know, talk basically. (Participant 4.2)
Translating	Test-taker seeking to understand by translating from L1 to the target language during a speaking or listening/watching activity	… the meaning in my mind is in Chinese, I think, not in English, but I know what she said. I need to write down something to make me remember the things, so I changed the Chinese word to English word in my mind. (Participant 2.3)
	Test-taker seeking to formulate speech by translating from L1 to the target language	I, I'll write my note taking combine using a English and my, my mother lounge language, so I'll, I'll little, feel little difficult to translate the my mother tongue language into English, so it takes so much time to look at the my note by using, explain by using a English. (Participant 2.6)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Inferencing	Test-taker seeking to understand by using information in the lecture to guess the meanings of linguistic items or to make up missing information	I found that the chimpanzee will use the stick for eating the insects, so I heard the, the word, "sharp," then I can determine the tool that chimpanzee will use the tool will very specific tool that, than the other stick, yup. (Participant 2.6)
Metacognitive strategies: involving organizing, planning, and evaluating		
Setting goals	Test-taker seeking a goal for completing a task	So but I should speak within 60 seconds, I should hurry up to say something, the main topics and the details, I should integrate all the things within 60 seconds, so in this case, I should control the speed of speaking more little bit faster than my normal speaking style. (Participant 1.1)
Identifying the purpose of the task	Test-taker identifying the purpose of the task: purposeful listening/watching and/or speaking	I just think, maybe I need to half of the time should use, should be used to say the two <unidentifiable> to say about the reliability and half of the time should, you should say the features... (Participant 3.2)
Planning	Test-taker planning the parts, sequence, or main ideas to be expressed verbally	I was like figure out, I was planning my speech. For that, I was focus on the to give two different definition of the tools. So I was happy about this. Okay, I figure out it. Okay, I have this. And I try to plan in my mind, you know ... And now I must take the plan what I want to say the 60 seconds, so I thought, okay, biology, the tools, the definitions, two definitions, narrow, broad, what they say about the this one and this one, okay, and examples, okay. (Participant 1.3)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Monitoring	Test-taker monitoring the clock while listening/watching, preparing, or speaking	I als-, I was also looking at the seconds, so I was real- I realized that there was not enough time for me to say kind of everything, so I man-, I managed myself to kind of keep the time... (Participant 4.3)
Self-correcting	Test-taker self-correcting errors in his/her own pronunciation, vocabulary, grammar, etc.	"Reliability is absence of problems and features," I said that, instead of "is", be verb. I tried to use "means." (Participant 1.1)
Evaluating the content of what was heard/seen	Test-taker evaluating the content of what he/she heard/saw	Yes, I just agree <laugh> his statement because in reality, actually, for example, the car, yeah, the reliability increase and stable between the car, but, so they, they change the design or extra conditions, features at the car, I think that, that is not necessary for the car, but they, at, at the design, the, I don't know <laugh> why they like a, how to say that, their, sometimes, the design is a little bit changed, not too much, but just some kind of add the line at the side of the car or they change the wheel more shiny… And then, like a, yeah, yeah, so I agreed with his statement. Yeah. (Participant 3.5)
Evaluating performance	Teat-taker evaluating language production while speaking	And yeah, I think so far, at this part, it was okay in terms of expressing <laugh> my opinion, and my kind of summary of the lecture. (Participant 4.3)
Evaluating language production	Test-taker evaluating language production after completing a task	I thought I didn't prepare and organized very well... (Participant 1.2)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Affective strategies: involving self-talk or mental control over affect		
Lowering anxiety	Test-taker reducing anxiety by taking a break or using techniques	I was trying to be very composed. (Participant 4.4)
Encouraging self	Test-taker encouraging him/herself through positive statements	I think, I, I, I told myself oh, this is, this is good sign <laugh> That means it's a familiar topic, that means I probably can perform well. (Participant 4.4)
Justifying performance	Test-taker justifying his/her performance	I was not doubting my, like you know, of course, I'm, I'm accented of, of the <laugh>, but the, the thing is I, I think, I think the thing is I, I, I think, I, I know the topic well, and then know how to present. I think I have the right information here if you give me like one, 30 mor-, 30 more seconds, I, I can give you a very good response, but then, if you cut off, you know, half of the thing, it's very awful, probably it's gonna be two, or three out of six, so I think it's not like, it's not reflecting my real speaking ability. (Participant 4.4)

adapted from Swain, Huang, Barkaoui, Brooks, and Lapkin, (2009), pp. 79-99

REFERENCES

Alderson, J. C. (1981). Report of the discussion on communicative language testing. In J. C. Alderson & A. Hughes (Eds.), *Issues in language testing* (ELT Documents No. 1111, pp. 55-65). London, UK: The British Council.

Bachman, L. F. (1990). Some persistent problems and future directions. In *Fundamental considerations in language testing* (pp. 296-359). Oxford, UK: Oxford University Press.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671-704.

Bachman, L. F., & Cohen, A. D. (1998). Language testing – SLA interfaces: An update. In *Interfaces between second language acquisition and language testing research* (pp. 1-31). New York, NY: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Bamford, J. (2004). Gestural and symbolic uses of the deictic *here* in academic lectures. In K. Aijmer & A.-B. Stenstrom (Eds.), *Discourse patterns in spoken and written corpora* (pp. 113-138). Amsterdam: John Benjamins.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series MS-19). Princeton, NJ: Educational Testing Service.

Bellés-Fortuño, B., & Campoy-Cubillo, M. C. (2010). 'I sort of feel like, um, I want to, agree with that for the most part …': Reporting intuitions and ideas in spoken academic discourse. In M. R. Campoy-Cubillo, B. Bellés-Fortuño, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 56-66). London, UK: Continuum.

Biber, D. (1988). *Variation across speech and writing*. New York, NY: Cambridge University Press.

Biber, D. (1994). An analytical framework for register studies. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 31-56). New York, NY: Oxford University Press.

Biber, D. (1995). *Dimensions of register variation*. New York, NY: Cambridge University Press.

Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. In P. Leistyna & C. F. Meyer, *Corpus analysis: Language structure and language use* (pp. 47-70). New York, NY: Rodopi.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes, 5*(2), 97-116.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36*(1), 9-48.

Biber, D., Conrad, S. M., Reppen, R., Byrd, H. P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus* (TOEFL Monograph Series MS-25). Princeton, NJ: Educational Testing Service.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientation and test-taker performance on English-for-Academic-Purposes speaking tasks* (TOEFL Monograph Series MS-29). Princeton, NJ: Educational Testing Service.

Brown, H. D. (2000). *Principles of language learning and teaching* (4th ed.). White Plains, NY: Longman.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph Series MS-20). Princeton, NJ: Educational Testing Service.

Camiciottoli, B. C. (2004). Interactive discourse structuring in L2 guest lectures: Some insights from a comparative corpus-based study. *Journal of English for Academic Purposes, 3*(1), 39-54.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Jr. Oller (Ed.), *Issues in language testing research* (pp. 333-342). Rowley, MA: Newbury House.

Canale, M. (1984). Testing in a communicative approach. In G. A. Jarvis (Ed.), *The challenge for excellence in foreign language education* (pp. 79-92). Middlebury, VT: The Northeast Conference Organization.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Carroll, B. J. (1980). *Testing communicative performance*. London, UK: Pergamon Institute of English.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). New York, NY: Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006). The technology thread. In *Assessing language through computer technology* (pp. 1-19). Cambridge, UK: Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language™*. New York, NY: Routledge.

Chapelle, C. A., & Lee, H-W. (2013). *What is argument-based validation?* Paper presented at the annual meeting of the Language Testing Research Colloquium, Seoul.

Charles, C., & Ventola, E. (2002). A multi-semiotic genre: the conference slide show. In E. Ventola, C. Shalom, & S. E. Thompson (Eds.), The language of conferencing. Frankfurt: Peter Lang.

Chaudron, C., & Richards, J. C. (1986). The effect of discourse markers on the comprehension of lectures. *Applied Linguistics, 7*(2), 113-127.

Collerson, J. (1994). *English grammar: A functional approach*. Newton, NSW: Primary English Teaching Association.

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System, 29*, 1-14.

Conrad, S., & Biber, D. (Eds.), (2001). *Variation in English: Multi-dimensional studies*. Harlow, England: Pearson Education.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly, 34*(2), 213-238.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2005). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL* (TOEFL Monograph Series MS-26). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (TOEFL Monograph Series MS-30). Princeton, NJ: Educational Testing Service.

DeCarrico, J., & Nattinger, J. R. (1988). Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes, 7*(2), 91-102.

Derewianka, B. (2001). Pedagogical grammars: Their role in English language teaching. In A. Burns & C. Coffin (Eds.), *Analysing English in a global context: A reader* (pp. 240-269). London, UK: Routledge.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations* (TOEFL Monograph Series MS-8). Princeton, NJ: Educational Testing Service.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, UK: Cambridge University Press.

Douglas, D. (2010). *Understanding language testing*. London, UK: Hodder Education.

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics, 27*, 115-132.

Droga, L., & Humphrey, S. (2002). Getting started with functional grammar. Berry, Australia: Target Texts.

Educational Testing Service. (2008). TOEFL® iBT Test Integrated Speaking Rubrics (Scoring Standards). Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/Integrated_Speaking_Rubrics_2008.pdf

Educational Testing Service. (2010). TOEFL iBT® Test Sample Questions. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf

Educational Testing Service. (2011). TOEFL iBT® Quick Prep: Volume 3. Retrieved from http://www.ets.org/s/toefl/pdf/qp_v3_web.pdf

Enikõ, C. (2000). Academic lectures: An interface of an oral and literate continuum. *NovELTy, 7*(3), 30-46.

Færch, C. & Kasper, G. (1983). Plans and strategies in foreign language communication. In *Strategies in interlanguage communication* (pp. 20-60). New York, NY: Longman.

Flowerdew, J., & Tauroza, S. (1995). The effect of discourse markers on second language lecture comprehension. *Studies in Second Language Acquisition, 17*(4), 435-458.

Fortanet, I. (2004). The use of 'we' in university lectures: reference and function. *English for Specific Purposes, 23*(1), 45-66.

Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243-265). Ann Arbor, MI: University of Michigan Press.

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345-369.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hillsdale, NJ: Lawrence Erlbaum.

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10*(1), 9-27.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*(2), 133-167.

Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT Journal, 15*(1), 85-88.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London, UK: Edward Arnold.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). New York, NY: Routledge.

Harrison, A. (1983). Communicative testing: Jam tomorrow? In A. Hughes & D. Porter (Eds.), *Current development in language testing* (pp. 77-85). London, UK: Academic Press.

Huang, H-T. D. (2010). Modeling the relationships among topical knowledge, anxiety, and integrated speaking test performance: A structural equation modeling approach. (Order No. 3417461, The University of Texas at Austin). *ProQuest Dissertations and Theses*. Retrieved from http://search.proquest.com/docview/748226412?accountid=10906. (748226412).

Hulstijn, J. H., Coopmans, P., Van Hout, R., & Bos, P. (2003). *Language acquisition and multilingualism.* Research programme of the Research Council for the Humanities of the Netherlands Organisation for Scientific Research (NWO).

Hymes, D. H. (1967). On communicative competence. Unpublished manuscript, University of Pennsylvania.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.). *Sociolinguistics* (pp. 269-293). London, UK: Penguin.

Inoue, M. (2009). Health Sciences Communication Skills Test: The development of a rating scale. *Melbourne Papers in Language Testing, 14*(1), 55-91.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (TOEFL Monograph Series MS-16). Princeton, NJ: Educational Testing Service.

Jamieson, J. M., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Framework for a new TOEFL. In C. A. Chapelle, M. K. Enright and J. M. Jamieson (Eds.) *Building a validity argument for the Test of English as a Foreign Language™* (pp. 55-96). New York, NY: Routledge.

Jin, T., & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing, 30*(1), 23-47.

Kane, M. (2012). Validating score interpretations and uses: Messick lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing, 29*(1), 3-17.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: Issues and practice, 18*(2), 5-17.2

Khuwaileh, A. A. (1999). The role of chunks, phrases, and body language in understanding co-ordinated academic lectures. System, 27(2), 249-260.

King, P. (1994). Visual and verbal messages in the engineering lecture: notetaking by postgraduate L2 students. In J. Flowerdew (Ed.), Academic listening: Research perspectives. Cambridge: Cambridge University Press.).

Lee, Y-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131-166.

Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing, 21*(3), 335-359.

Levie, W. H. (1987). Research on pictures: A guide to the literature. In D. M. Willows & H. A. Houghton (Eds.), *The psychology of illustration: Vol. 1. Basic research* (pp. 1-50). New York: Springer-Velag.

Levis, J., Levis, G.M., & Slater, T. (2012). Written English into spoken: A functional discourse analysis of American, Indian, and Chinese TA presentations. In G. Gorsuch

(Ed.), *Working theories for TA and ITA development* (pp. 529-572). Stillwater, OK: New Forums Press.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing, 17*(1), 43-64.

Lindemann, S., & Mauranen, A. (2001). "It's just real messy": The occurrence and function of *just* in a corpus of academic speech. *English for Specific Purposes, 20*(Supplement 1), 459-475.

Liu, H. (2005). *An investigation of methods for assessing authenticity in computer-assisted language learning and assessment*. Unpublished master's thesis, Iowa State University.

Londe, Z. C. (2009). The effects of video media in English as a second language listening comprehension tests. *Issues in Applied Linguistics, 17*(1), 41-50.

Low, M. (2010). Teachers and texts: Judging what English language learners know from what they say. In A. Paran & L. Sercu (Eds.), *Testing the untestable in language education* (pp. 241-255). Bristol, UK: Multilingual Matters.

Lumley, T., & Brown, A. (1998). An investigation into the authenticity of discourse in a specific-purpose language performance test and its relevance to test validity. In E. L. Siu-leung & G. James (Eds.), *Testing and evaluation in second language education* (pp. 22-33). Hong Kong, China: Language Centre.

Mauranen, A. (2001). Reflexive academic talk: Observations from MICASE. In R. C. Simpson & J. M. Swales (Eds.), *Corpus linguistics in North America: Selections from the 1999 symposium* (pp. 165-178). Ann Arbor, MI: The University of Michigan Press.

Mauranen, A. (2003). "But here's a flawed argument": Socialisation into and through metadiscourse. In P. Leistyna & C. F. Meyer (Eds.), *Corpus analysis: Language structure and language use* (pp. 19-34). New York, NY: Rodopi.

Mauranen, A. (2004). "They're a little bit different"… : Observations on hedges in academic talk. In K. Aijmer & A.-B. Stenström (Eds.), *Discourse patterns in spoken and written corpora* (pp. 173-197). Amsterdam: John Benjamins.

Miller, H. B., & Burton, J. K. (1994). Images and imagery theory. In D. M. Moore & F. M. Dwyer (Eds.), *Visual literacy* (pp. 65-83). Englewood Cliffs, NJ: 298 Educational Technology Publications.

Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment*. (CRESST Report 800). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., & Haertel, G. D. (2006a). *Implications of evidence-centered design for educational testing*. (PADI Technical Report 17). Menlo Park, CA: SRI International.

Mislevy, R. J., & Haertel, G. D. (2006b). Implications of evidence-centered design for educational testing. *Educational measurement: Issues and practice, 25*(4), 6-20.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K., & John, M. (2014). *Psychometric considerations in game-based assessment*. (White Paper). Redwood City, CA: GlassLab, Inc.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives, 1*(1), 3-62.

Mohan, B., & Slater, T. (2005). A functional perspective on the critical 'theory/practice' relation in teaching language and science. *Linguistics and Education, 16*(2005), 151-182.

Mohan, B., Leung, C., & Slater, T. (2010). Assessing language and content: A functional perspective. In A. Paran & L. Sercu (Eds.), *Testing the untestable in language education* (pp. 217-240). Bristol, UK: Multilingual Matters.

Moore, N.-J., Hickson, III, Mark, & Stacks, D. W. (2009). *Nonverbal communication: Studies and applications* (5th ed.). Oxford, UK: Oxford University Press.

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143-157). Oxford, UK: Oxford University Press.

Nattinger, J. (1986). Lexical phrases, functions and vocabulary acquisition. *The ORTESOL Journal, 7*, 1-14.

Nattinger, J. R., & DeCarrico, J. S. (1992). Teaching spoken discourse: Listening comprehension. In *Lexical phrases and language teaching* (pp. 131-156). Oxford, UK: Oxford University Press.

Ogborn, J., Kress, G., Martins, I., & McGillicuddy, K. (1996). *Explaining science in the classroom*. Buckingham, UK: Open University Press.

Ohkubo, N. (2009). Validating the integrated writing task of the TOEFL internet-based test (iBT): Linguistic analysis of test takers' use of input material. *Melbourne Papers in Language Testing, 14*(1), 1-31.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*(1), 26-56.

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25*(2), 237-272.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26*(4), 561-587.

Plakans, L. (2013). In *The encyclopedia of applied linguistics* (Vol. 1). Hoboken, NJ: Wiley-Blackwell.

Plakans, L. M., & Gebril, A. (2012). A close investigation into source use in L2 integrated writing tasks. *Assessing Writing, 17*(1), 18-34.

Progosh, D. (1996). Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal, 14*(1), 34-44.

Ravelli, L. (2000). Getting started with functional analysis of texts. In L. Unsworth (Ed.), *Researching language in schools and communities: Functional linguistic perspectives* (pp. 27-64). London, UK: Cassell.

Raymond, M., & Neustel, S. (2006). Determining test content of credentialing examinations. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 181-224). Mahwah, NJ: Erlbaum

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1-32.

Reppen, R. (2004). Academic language: An exploration of university classroom and textbook language. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions: Perspective from corpus linguistics* (pp. 65-86). Amsterdam: John Benjamins.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels*. (TOEFL Monograph Series MS-21). Princeton, NJ: Educational Testing Service.

Rosenfeld, M., Oltman, P. K., & Sheppard, K. (2004). *Investigating the validity of TOEFL: A feasibility study using content and criterion-related strategies*. (TOEFL Research Reports 71). Princeton, NJ: Educational Testing Service.

Royce, T. D. (2007). Intersemiotic complementarity: A framework for multimodal discourse analysis. In T. D. Royce & W. L. Bowcher (Eds.), *New directions in the analysis of multimodal discourse* (pp. 63-109). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 5-30.

Schleef, E. (2008). The "lecturer's OK" revisited: Changing discourse conventions and the influence of academic division. *American Speech, 83*(1), 62-84.

Simpson, R. C. (2004). Stylistic features of academic speech: The role of formulaic expressions. In U. Connor & T. A. Upton (Eds.). *Discourse in the professions: Perspectives from corpus linguistics* (pp. 37-64). Amsterdam: John Benjamins.

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan corpus of academic spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 38*(3), 419-441.

Simpson-Vlach, R. & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512.

Slater, T., & Butler, J. I. (2015). Examining connections between the physical and the mental in education: A linguistic analysis of PE teaching and learning. *Linguistics and Education, 30*(2015), 12-25.

Slater, T., Levis, G., & Levis, J. (2015). Spoken parentheticals in TA and ITA instructional discourse in Social Science and STEM disciplines: The interaction of intonational, ideational, and interpersonal resources in signaling information structure. In G. Gorsuch (Ed.), *Talking matters: Research on talk and communication of international teaching assistants*. Stillwater, OK: New Forums Press.

Slater, T. & Mohan, B. (2010). Towards systematic and sustained formative assessment of causal explanations in oral interactions. In A. Paran & L. Sercu (Eds.), *Testing the untestable in language education* (pp. 256-269). Bristol, UK: Multilingual Matters.

Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing, 18*(4), 463-481.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2*(1), 31-40.

Steinberg, L. S., Mislevy, R. J., Almond, R. G., Baird, A. B., Cahallan, C., Divello, L. V., Senturk, D., Yan, D., Chernick, H., & Kindfield, A. C. H. (2003). *Introduction to the Biomass project: An illustration of evidence-centered assessment design and delivery capability*. (CSE Report 609). Los Angeles, CA: University of California, Center for the Study of Evaluation (CSE).

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York, NY: Cambridge University Press.

Sternberg, R. J. (1988). *The triarchic mind: A new theory of human intelligence*. New York, NY: Viking.

Strodt-Lopez, B. (1991). Tying it all in: Asides in university lectures. *Applied Linguistics, 12*(2), 117-140.

Sueyoshi, A., & Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661-699.

Suvorov, R. & Hegelheimer, V. (2013). Computer-assisted language testing. In A. J. Kunnan (Ed.), *The companion to language assessment, vol 2: Approaches and development* (pp. 594-614). West Sussex, UK: Wiley-Blackwell.

Swain, M., Huang, L-S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT™ (SSTiBT): Test-takers' reported strategic behaviors*. (TOEFL iBT™ Research Report TOEFLiBT-10). Princeton, NJ: Educational Testing Service.

Swales, J. M. (2001). Metatalk in American academic talk: The cases of *point* and *thing*. *Journal of English Linguistics, 29*(1), 34-54.

Swales, J. M., & Burke, A. (2003). "It's really fascinating work": Differences in evaluative adjectives across academic registers. In P. Leistyna & C. F. Meyer (Eds.), *Corpus analysis: Language structure and language use* (pp. 1-18). New York, NY: Rodopi.

Swales, J. M., & Malczewski, B. (2001). Discourse management and new-episode flags in MICASE. In R. C. Simpson & J. M. Swales (Eds.), *Corpus linguistics in North America: Selections from the 1999 symposium* (pp. 145-164). Ann Arbor, MI: The University of Michigan Press.

Thompson, S. E. (2003). Text-structuring metadiscourse, intonation and the signaling of organisation in academic lectures. *Journal of English for Academic Purposes, 2*(1), 5-20.

Wagner, E. (2006). *Utilizing the visual channel: An investigation of the use of video texts on tests of second language listening ability*. (Unpublished doctoral dissertation). Teachers College, Columbia University, New York, NY.

Wagner, E. (2010a). Test-takers' interaction with an L2 video listening test. *System, 38*, 280-291.

Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing, 27*(4), 493-513.

Webb, N. (2006). Identifying content for student achievement tests. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Erlbaum.

Weir, C. J. (1990). Test methods. In *Communicative language testing* (pp. 42-85). New York: Prentice Hall International.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford, UK: Oxford University Press.

Widdowson, H. G. (1979). *Explorations in applied linguistics*. Oxford, UK: Oxford University Press.

Widdowson, H. G. (1983). *Learning purpose and Language use*. Oxford, UK: Oxford University Press.

Wu, W. M., & Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language Testing, 18*(2), 187-206.

Yu, G. (2013). From integrative to integrated language assessment: Are we there yet? *Language Assessment Quarterly, 10*(1), 110-11